COMBINING MULTIVARIATE STATISTICAL METHODS AND SPATIAL ANALYSIS TO CHARACTERIZE WATER QUALITY CONDITIONS IN THE WHITE RIVER BASIN, INDIANA, U.S.A.

Andrew Stephan Gamble

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science in the Department of Earth Science,
Indiana University

September 2010

Accepted by the Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Meghna Babbar-Sebens, Ph. D., Chair

Lenore P. Tedesco, Ph. D.

Master's Thesis
Committee

Hanxiang Peng, Ph. D.

ACKNOWLEDGEMENTS

I would like to thank my friends and family who were always there for me as I pursued this degree. I do not know how I would have accomplished this without you. I would also like to thank my advisor, Dr. Meghna Babbar-Sebens, for taking taking me on as her first graduate student at IUPUI. I will be forever grateful for the opportunity that she gave me, right before I was about to change my focus from Earth Science. Thank you to Dr. Lenore Tedesco for your help in keeping this research focused on reality. Thank you to Dr. Hanxiang Peng for your guidance on statistical methods. I would like to thank the Indiana Department of Environmental Management and the Center for Earth and Environmental Science for providing the data used in this study. Lastly I would like to thank the Indiana State Department of Agriculture for providing funding for this research.

ABSTRACT

Andrew Stephan Gamble

COMBINING MULTIVARIATE STATISTICAL METHODS AND SPATIAL ANALYSIS TO CHARACTERIZE WATER QUALITY CONDITIONS IN THE WHITE RIVER BASIN, INDIANA, U.S.A.

This research performs a comparative study of techniques for combining spatial data and multivariate statistical methods for characterizing water quality conditions in a river basin. The study has been performed on the White River basin in central Indiana, and uses sixteen physical and chemical water quality parameters collected from 44 different monitoring sites, along with various spatial data related to land use – land cover, soil characteristics, terrain characteristics, eco-regions, etc. Various parameters related to the spatial data were analyzed using ArcHydro tools and were included in the multivariate analysis methods for the purpose of creating classification equations that relate spatial and spatio-temporal attributes of the watershed to water quality data at monitoring stations. The study compares the use of various statistical estimates (mean, geometric mean, trimmed mean, and median) of monitored water quality variables to represent annual and seasonal water quality conditions. The relationship between these estimates and the spatial data is then modeled via linear and non-linear multivariate methods. The linear statistical multivariate method uses a combination of principal component analysis, cluster analysis, and discriminant analysis, whereas the non-linear multivariate method uses a combination of Kohonen Self-Organizing Maps, Cluster

Analysis, and Support Vector Machines. The final models were tested with recent and independent data collected from stations in the Eagle Creek watershed, within the White River basin. In 6 out of 20 models the Support Vector Machine more accurately classified the Eagle Creek stations, and in 2 out of 20 models the Linear Discriminant Analysis model achieved better results. Neither the linear or non-linear models had an apparent advantage for the remaining 12 models. This research provides an insight into the variability and uncertainty in the interpretation of the various statistical estimates and statistical models, when water quality monitoring data is combined with spatial data for characterizing general spatial and spatio-temporal trends.

Meghna Babbar-Sebens, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

GIS and remote sensing technology create means to measure various spatial characteristics – e.g., land cover, geomorphologic, climatic, geologic, hydrologic, and ecologic parameters - associated with non-point pollution sources in river basins (Ward and Trimble, 2004). Quantitative assessment of these non-point pollution sources is needed, in order to better manage the relationship between human impact on the land and water quality. Additionally, non-point source pollution such as combined sewer overflows, have a great effect on water quality, especially during low flow periods (Fenelon, 1998). Anthropogenic sources of pollution greatly affect the water quality in agricultural and urban areas. For example, runoff from row crop agriculture has resulted in excess fertilizer in the White River watershed. This has resulted in an elevated nutrient level that has caused problems in the tributaries and reservoirs (e.g. excess eutrophication) that make up the White River watershed and is a leading cause of eutrophication in the Gulf of Mexico (Goolsby et al., 2000). Urban sources can also have an impact on water quality in a river basin. For example, industrial and wastewater treatment discharges and road runoff can greatly increase the salinity of surrounding water bodies, as well as introduce other toxic substances, metals, and pharmaceuticals. To add to the complexity, changes in the landscape throughout the watershed have led to significant temporal changes in the nature and contaminant loadings of various non-point sources of pollution. Regular monitoring can alleviate some of these challenges and help identify the contaminant sources and trends in water quality conditions (USEPA, 2007). However, regular and spatially rigorous monitoring can be expensive and, therefore, limit the number of monitoring sites and the frequency of monitoring in a river basin. For this reason, a screening method can be useful in characterizing water quality in the unmonitored tributaries of a river basin and for analyzing the conditions and impact of land-use over time on water quality.

Several studies have developed empirical models that can be used to predict water quality. Linear multivariate approaches that combine principal component analysis (PCA)/factor analysis (FA), cluster analysis (CA), and linear discriminant analysis (LDA) have been used in many water quality prediction studies (Santos-Roman et al., 2003; Paul et al., 2006; Jenerette et al., 2002; Snelder et al., 2005; Iscen et al., 2007; Frohlich et al.,

2007). Santos-Roman et al. (2003) used a combination of FA, CA, and LDA methods to predict water quality in unmonitored watersheds in Puerto Rico. A FA was used to reduce the number of physical and chemical parameters into fewer variables. Using parameters determined by the FA, a CA grouped the watersheds into five clusters: forested, urban-polluted, mixed urban/forested, plutonic forested, and limestone. Each cluster's water quality was described based on the mean value of the chemical constituents selected in the factor analysis. A LDA using physical attributes of each watershed was then performed to predict membership into one of the five clusters. The physical attributes used were: rate of change of forest land cover from 1977 to 1991/1998, percentage of limestone, mean annual rainfall, and shape factor. The rate of change of forest land cover was most successful in discriminating between clusters. Prediction equations, derived from the LDA, were formulated that allow for a user to insert the aforementioned physical attributes of an unmonitored watershed and determine to which water quality cluster that watershed belongs.

Paul et al. (2006) used similar techniques that clustered watersheds based on related watershed characteristics. The goal of this study was to look at fecal coliform data and group impaired streams based on point and non point sources in each streams' watershed. Snelder et al. (2005) used PCA and CA to show the classification strength of an existing mapped classification of rivers in New Zealand. Iscen et al. (2007) used PCA/FA and CA to classify water quality at twelve different sites in Uluabat Lake, Turkey. Frohlich et al. (2007) found that lithologic signals and anthropogenic point sources caused differences in stream chemistry in the Dill River watershed in Germany using PCA/FA and CA. Snapshot data at low flow, high flow, and mean flow, rather long term historical data, was used in the Frohlich et al. study.

All of these studies have certain limitations to their methods. The Santos-Roman et al. (2003) study had issues with limited data because samples were taken only a few times per year for 23 years. Paul et al. (2006) also had limitations due to data availability. The Iscen et al. (2007), Santos-Roman et al. (2003), and Frohlich et al. (2007) studies used only the mean values of water quality variables while conducting the PCA and CA. Santos-Roman et al. (2006) considered the median, but this study did not show a difference between clustering the mean versus clustering median time averaged

data. This may have been a result of the lack of data available for their study. Lastly, the methodology, in all of the aforementioned studies, was limited by the assumption of statistical linearity.

An increasingly popular approach to the clustering and classification of data is the use of nonlinear empirical modeling techniques, such as artificial neural networks (ANN) and support vector machines (SVM). The main advantage of these non-linear techniques to the linear multivariate techniques is that they can learn the non-linear dependencies between variables in a complex system, without the knowledge of the underlying processes. For example, for simulation of dependencies between various drivers and their effects in a watershed, these methods do not require specific information about the underlying hydrological sub processes to create a model (Jiang and Nan, 2006). Application of artificial neural networks and support vector machines to the environmental field, and, specifically, in the prediction of water quality has been explored in multiple studies (e.g., Bowers and Shedrow, 2000; Park, 2003; Yunrong and Liangzhong, 2009). The European Commission conducted a study called PAEQANN that used artificial neural networks to provide a predictive tool that would better enable lawmakers to enact effective policies in freshwater management (Park, 2003). In this study, the PAEQANN researchers used a type of ANN, the Kohonen Self-Organizing Map (SOM), to form ecology-based regionalization. They applied the SOM to data that described the presence or absence of diatom species, and derived clusters based on the results. In another study, Bowers and Shedrow used another type of ANN, the Back Propagation ANN (BP-ANN), to create a predictor model of water quality. They selected precipitation, flow rate, and turbidity as input variables in order to predict suspended solids using a BP-ANN at their Savannah River, Georgia site. A different study by Yunrong and Lianzhong (2009) compared the performance of a SVM and a BP-ANN in the prediction of certain water quality variables. In their study they used ten different water quality variables to predict the future values of Chemical Oxygen Demand and Dissolved Oxygen. They concluded that the SVM outperformed the BP-ANN in terms of model forecasting accuracy.

The choice of linear and non-linear statistical approaches for designing empirical models is a key aspect in the current study. Linear and non-linear multivariate techniques

have both been shown to be effective water quality prediction techniques. A comparative analysis of both linear and nonlinear techniques can provide greater insight into the study of forecasting river water quality conditions. In this study, the PCA + CA + LDA methodology, as described by Santos-Roman et al. (2003) was applied to the White River. A parallel non-linear methodology that also used physical watershed variables to predict water quality conditions was proposed and tested. This methodology applied a SOM – CA methodology (similar to the PAEQANN study) to create water quality clusters. Then these results were combined with an empirical classification model created by an SVM using physical watershed variables as inputs. Additionally, long term water quality data was time averaged and used in conjunction with the physical watershed data. The overall objective of this research was to evaluate existing classification methods used for the screening of water quality conditions in the White River watershed. The methods are tested for the White River basin in Indiana based on the following specific objectives:

- Compare statistical multivariate models that use spatial and temporal characteristics to predict water quality conditions at unmonitored sites in the White River basin based on: (1) the choice of statistical indicator (i.e. mean, median, trimmed mean, and geometric mean) for time averaging water quality data, (2) the choice of time averaging based on seasonal or annual durations, and (3) the choice of using a linear or non-linear methodology
- Validate the models using water quality monitoring data not in the original data set.

MATERIALS AND METHODS

Case Study – White River Watershed

The White River Basin drains 11,350 square miles of central and southern Indiana and is part of the Mississippi River system (Jacques and Crawford, 1991). Stream flow in the watershed generally peaks in the spring months and is lowest in the late summer and fall (Fenelon, 1998). The entire basin can be divided into eight different sub-watersheds that have 8-digit hydrologic unit codes (HUC). They are the Upper White, Lower White, Eel, Driftwood, Flatrock-Haw, Upper East Fork White, Muscatatuck, and Lower East Fork (Figure 1). Agriculture accounts for about 70% of the land use throughout the basin, with most of the crop production coming from rotational soybeans and corn. Urban land use makes up approximately 8% of the watershed, and, as of 1990, 2.1 million people live in the entire basin. However, three-fourths of the population in the basin is located in Upper White, which contains the largest metropolitan areas of Indianapolis, Anderson, and Muncie. These three cities represent a significant amount of industrial development. The south-central portion of the basin is not as extensively farmed since it is unglaciated, has poor soils, and is much hillier. Most of the forested landscape is located in this area, which makes up approximately 22% of the watershed. Significant uses of the surface water withdrawn from the White River include thermoelectric power, industrial and mining uses, irrigation and livestock, and public drinking water supply (Fenelon, 1998).

For this study, water quality data was collected from the Indiana Department of Environmental Management (IDEM) fixed station database. The 2 main branches in this watershed are the main branch of the White River and the East Fork of the White River. Respectively, the White River main branch and the East Fork of the White River have 11 and 5 water quality monitoring stations located directly on them. There are 2 monitoring stations that are located downstream from the junction of these branches, and the remaining 26 monitoring stations are located on tributaries feeding these two main branches (Figure 2). The IDEM fixed station database is historic and ranges from 1991 to 2008 for the current study – data collection is ongoing. Water quality samples are generally taken monthly for these stations. From this database, a combination of 44 stations and 16 water quality variables met the requirements of completeness to prepare

the dataset for this study. Physical watershed attribute data was obtained (explained in detail below in a later section) by delineating the watersheds of interest in a geographic information system (GIS). Spatial data for the White River watershed is extensive and freely available from a variety of internet databases.

Water Quality Data Preparation

Indiana Department of Environmental Management Fixed Station Monitoring Database

Before any multivariate statistics were run, the water quality data from the Indiana Department of Environmental Management (IDEM) fixed station monitoring database had to be sorted and prepared. The goal in data preparation was to create an *n* x *m* data matrix, with *n* representing water quality monitoring stations and *m* representing water quality variables. The first step in accomplishing this process was to determine what combination of stations and variables would be acceptable for this study. Originally, 46 stations and 17 water quality variables were considered because of data availability. Data quality and outliers would later reduce the size of this dataset. However, before this reduction occurred, the datasets were divided into an annual dataset and four quarterly datasets. The quarterly datasets were defined as January 1 – March 31 (Quarter 1), April 1 – June 30 (Quarter 2), July 1 – September 30 (Quarter 3), and October 1 – December 31 (Quarter 4). These time periods were chosen to represent seasonal changes in water quality. Additionally these time periods can be used to reflect the different flow regimes of the watershed, with higher flows expected in Quarters 1 and 2 and lower flows expected in Quarters 3 and 4 (Fenelon, 1998). Four different statistical indicators were chosen to time-average each water quality variable at each site: mean, median, trimmed mean, and geometric mean. Different statistical indicators were taken in order to determine if they caused differences in the clustering or classification to be conducted in the multivariate analysis. The apparent advantage of the mean lies in the fact that it contains all of the information about all of the data; however, this can also be a disadvantage when large outliers skew the true value of a data point. Thus, the apparent advantage to the median and geometric mean can be attributed to their robustness to outliers. The trimmed mean is considered semi-robust since it removes the largest and smallest values (for this study 5% of the data at each extreme was removed), and takes the mean of the remaining data.

Accounting for the values of observations below the detection limit was an issue in calculating the different statistical indicators. The regression on order statistics (ROS) method was used to estimate the value for the missing observation (Singh et al., 2006). The regression methods are parametric in nature and assume a normal, log normal, or gamma distribution. Essentially, the slope and intercept of a regression line are computed using detected data, and the non-detect data is estimated by this regression line (Singh et al., 2006). The recommended ROS method for environmental data is known as the Helsel's robust ROS, and it is performed by extrapolating the non-detect data in log scale, then transforming the results back to the original scale (Singh et al., 2006). The statistical program proUCL 4.0 (Singh et al., 2007) was used to estimate the non-detectable data, and after the ROS method was complete, the modified datasets were used to calculate statistical indicators – means, trimmed means, and geometric means. The ProUCL 4.0 software was developed to estimate the upper confidence limit (UCL) of an unknown population mean, and it includes other statistical tools, such as the ROS tool. After combining the different annual and quarterly datasets and 4 different statistical indicators, 20 data matrices were formed.

The newly created datasets were investigated for potential problems arising from data quality and outliers among the monitoring stations. The first issue of data quality arose with the *Escherichia coli (E. coli)* data. This water quality variable differed from the other water quality parameters that were chosen because it was not as frequently sampled as the other parameters. Additionally, *E. coli* values are highly dependent on the timing and location of a sample and therefore highly variable. Lastly, the methodology in the IDEM dataset for determining *E. coli* changed in 1999 from colony forming units/ 100 ml to most probable number/100 ml. After considering the few samples of *E.coli* at each station, the lack of reliable sampling, and the change in methodology in 1999, it was determined that *E. coli* would not be a practical parameter to describe water quality conditions for this study. No data quality issues were found with the remaining 16 variables, and these were the variables chosen to give a general description of the water quality conditions at each given site. Table 1 shows the 16 chosen variables. The second issue dealt with in constructing the final data set was identifying very large outliers that could cause problems with the future analyses. Two stations had abnormally large values

of certain variables, such as alkalinity and specific conductance. After further investigation it was determined that these stations monitored underground rivers. Since this study is investigating surface water quality, these two stations were removed permanently. The final dataset was composed of a 44 stations x 16 water quality variable matrix.

Eagle Creek Watershed Management Plan Database

The Eagle Creek Watershed Management Plan (ECWMP) (Tedesco et al., 2005) database was used to test the performance of the models made from the IDEM water quality data. The ECWMP datasets were prepared exactly the same way as the IDEM data with a few key differences. In this dataset 11 sites were sampled from March 2007 to March 2010 for the current study – dataset is ongoing. Water quality variables were time averaged with the same statistical indicators for an annual dataset and quarterly datasets, and the non-detectable data was estimated using the ROS method. However, some of the water quality variables were missing or prepared differently in the ECWMP dataset. Chemical Oxygen Demand and Total Iron were not sampled in the ECWMP and could therefore not be included in the dataset. Additionally, nitrate and nitrite were measured as separate variables in the ECWMP dataset, so they were simply added together to make them comparable to the IDEM dataset.

Watershed Delineation

The ArcHydro toolbox and a 30-meter digital elevation model (DEM) of the White River Watershed were used for delineating the watershed drainage area of each water quality monitoring station (ESRI, 2005). Before delineation could take place, the raw DEM had to be preprocessed and several additional grids were created. The AGREE method, developed at the University of Texas at Austin in 1997, was used to recondition the DEM for watershed delineation (Hellweger, 1997). The White River watershed stream network, as described by the National Hydrography Dataset (NHD), was first "burned" into the DEM. This ensured that the stream network derived from the DEM is close to reality. Additionally, any sinks or depressions in the DEM were filled, so the delineation algorithm did not create false watersheds. After these two steps were complete, a flow direction grid was created. This grid shows the direction water will flow by indicating the direction of steepest descent from one cell to another. The next

grid created was the flow accumulation grid. This grid uses the flow direction grid to determine the number of cells upstream of a given cell, and can be used to define the stream grid. With the stream grid defined, the stream is then broken up into segments, and catchments are defined for each of these stream segments. At this point the locations of the water quality monitoring stations are located and the watersheds for each station are defined.

Variable Reduction

Linear – Principal Component Analysis

Principal component analysis (PCA) is a variable reduction procedure used when dealing with a large number of variables believed to be correlated with each other (Suhr, 2005). Redundant variables are reduced to artificial variables called principal components or factors which account for most of the variance in the data and are orthogonal (and, therefore linearly independent) to each other. Deriving principal components is accomplished by finding the eigenvalues of the covariance matrix of the original variables. The PCA model is:

$$\boldsymbol{Y} = \boldsymbol{XB}, \qquad (1)$$

where $\boldsymbol{Y}$ is a matrix of observed input variables, $\boldsymbol{X}$ is a matrix of factor scores, and $\boldsymbol{B}$ is a matrix of eigenvectors or the factor pattern.

Since variables are not necessarily scaled the same, they are standardized so that they are comparable (Fodor, 2002) Once factors are calculated, it is necessary to determine the number of meaningful components to retain. There are four commonly used approaches to determine this: minimum eigenvalue equals one method/Kaiser criterion, Scree test, proportion of variance accounted for, and the interpretability criteria (Suhr, 2005). In this study, the Kaiser method was used, which retains any factor whose eigenvalue is greater than one. The reasoning for this is that an eigenvalue of one would be the amount of variance accounted for by one variable, and any eigenvalue greater than one explains more variance due to additional variables (SAS, 2002-2004). Additionally, varimax rotation was also used, so that high variable loadings are easily recognizable (SAS, 2002-2004). The varimax rotation involves maximizing the variance of the loadings of each factor (Davis, 2002). Factor variance is defined by:

$$s_k^2 = \frac{p\sum_{j=1}^{m}\left(a_{jp}^2/h_j^2\right)^2 - \left(\sum_{j=1}^{m} a_{jp}^2/h_j^2\right)^2}{p^2}, \quad (2)$$

where p is the number of factors, m is the number of original variables, a_{jp} is the loading of variable j on factor p, and h_j^2 is the communality of the jth variable. Additionally, varimax rotation searches iteratively for a linear combination of factors, such that variance is maximized by:

$$\max(V) = \sum_{k=1}^{p} s_k^2 \quad (3)$$

Non-linear – Kohonen Self-Organizing Map

The Kohonen self-organizing map (SOM) is an unsupervised artificial neural network (ANN) made up of two layers, inputs and outputs that projects multidimensional inputs onto 2-dimensional (in this case) space. The map or grid is made up of a user defined topology and number of neurons (Rojas, 1996). The neurons are given weights which are initialized randomly. Figure 3 shows the architecture of a simplified SOM. In addition, a learning constant and neighborhood function are selected (Rojas, 1996). At this point the SOM is ready to be trained. In each of the iterations of the training, an input vector is chosen randomly and Euclidean distance is calculated between the input vector and all the weight vectors in the map. Euclidean distance is calculated by:

$$Dist = \sqrt{\sum_{i=0}^{i=n}(V_i - W_i)^2} \quad (4)$$

where V_i is the input vector and W_i is the weight vector.

The most similar neuron to a given input vector, or best matching unit (BMU), and the weight vectors of the neurons around this unit are adjusted to be closer to the input vector. During the training process the neighborhood radius and learning rate are decreased over time (Vesanto et al., 2000). Training usually occurs in two phases: rough training and fine-tuning. In the rough training phase the neighborhood radius and learning are relatively large, and the map takes its basic form. In the fine-tuning phase neighborhood radius and learning rate initialize at much smaller values (Vesanto et al., 2000). After the SOM is trained the Euclidean distance between nodes can be examined in a visualization grid known as the unified distance matrix, which can be very useful with clustering data.

Data Transformations

Two important assumptions for the PCA are that variables are normally distributed and the measurement scale is interval or ratio type (Suhr, 2005). Box-Cox transformations are a common way to transform a set of variables to making them linear (Box and Cox, 1964). The Box-Cox transformation's most common form is:

$$y_i^{(\lambda)} = \frac{y_i^{\lambda}-1}{\lambda}, \text{ if } \lambda \neq 0; \; y_i^{(\lambda)} = logy_i, \text{ if } \lambda = 0 \qquad (5)$$

where y is the variable being transformed and λ is the power transformation.
The parameter λ is determined through maximum likelihood estimation of the likelihood function (Kutner et al., 2004). Basically, this parameter is used to transform a given variable, so that it is closest to normal as possible. The Shapiro-Wilk goodness of fit test can then be applied to the transformed data to indicate if the data is not normal with a certain level confidence. This test is designed for datasets with sample sizes between 3 and 5000 (Hammer et al., 2009).

In addition to the normality assumption, PCA assumes that data is in interval or measurement scale, so a standardization transformation was necessary (Suhr, 2005). Standardizing variables is also recommended when constructing the Kohonen SOM since the map is based on Euclidean distances and data on larger scales will dominate map organization (Vestano et al., 2000). The softmax transformation was chosen for preprocessing in the analyses. The equation for softmax scaling is shown below in two steps (Collica, R.S.):

$$x = \frac{(v-v_{mean})}{\lambda\left(\frac{\sigma_v}{2\pi}\right)}, \; s = \frac{1}{1+e^{-x}} \qquad (6)$$

where v is the variable to be scaled, σ_v is the standard deviation, and λ is the linear response to standard deviations.
The second part of the equation is referred to as the logistic function, and the first part scales the linear portion of the logistic function. This transformation is more or less linear in the middle range of values, and it has a smooth nonlinearity at both ends which ensures all values are in the [0 1] range and dampens the effect of outliers (Vestano et al., 2000). This standardization technique was used in all instances that data needed to be standardized.

Clustering Methodology

K-Means Clustering

Cluster analysis is a method that was used to assemble the output of the PCA and the SOM into homogeneous groups, where members are distinct to their group only (Davis, 2002). *K*-means clustering was used in this work to cluster monitoring stations with similar water quality characteristics. *K*-means cluster analysis is a divisive clustering method with *k* number of groups set *a priori* to analysis (Akume and Weber, 2002). The goal of the *K*-means method is to minimize the function of f_Σ for a given number of clusters (Akume and Weber, 2002). Each cluster has a centroid $\hat{z}_j$, which is defined as the mean value vector of the elements in its cluster C_j. The minimization equation is given as:

$$f_\Sigma(C) = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} \|x_i - \hat{z}_k\|_2^2 \quad (7)$$

Additionally, the cluster centroid of each cluster C_j is calculated as:

$$\hat{z}_{ji} := \frac{1}{|C_k|} \sum_{l=1}^{|C_k|} x_{li}, \text{ for } i=1, \ldots, m. \quad (8)$$

Once the number of clusters is set and cluster centroids are initialized, observations are added iteratively to the most similar cluster, whose centroid is then recalculated until all of the observations are grouped (Davis, 2002). The drawback to this method is that is difficult to effectively initialize values for the cluster centroids, so that the optimal clustering arrangement is formed. Therefore, several iterations of the clustering algorithm are run to ensure an optimal clustering arrangement is achieved (Akume and Weber, 2002). A two-level clustering approach was applied in this study by applying the *K*-means clustering method to the first four most important factors from the PCA, and to each SOM that was produced. The alternative to the two-level approach would be to cluster the raw water quality data. The main benefit of clustering the stations after variable reduction, rather than clustering the actual data is the reduction in computational cost. Even with a relatively small sample size, clustering algorithms can become extremely complex (Vestano and Alhoniemi, 2000).

Cluster Identification

The Davies-Bouldin cluster validity index was used to help determine the most correct number of clusters in the dataset. Since the appropriate number of clusters is not

known *a priori* to the analysis, several analyses for different numbers of clusters, *k*, were run to determine the most likely number of clusters. Also, the initialization of the clustering algorithm is random, so several iterations of the *k*-means algorithm are needed until convergence of the same cluster arrangement is reached at each level of *k* (Bezdek and Pal, 1998). The Davies-Bouldin index can then be examined at each level of *k* in order to identify the most likely number of clusters. The ratio of cluster scatter within the *ith* cluster and the separation between the *ith* and *jth* cluster defines this index. The within cluster scatter is defined by:

$$S_{i,q} = \left(\frac{1}{|X_i|}\sum_{x\epsilon X_i}\|x - \hat{z}_i\|_2^q\right)^{1/q}, \quad (9)$$

where z_i is the cluster centroid, x is a vector of the sample observations. Additionally, for a given cluster U, $\hat{z}_i$ is the cluster centroid, and, with within cluster scatter defined, between cluster separation is defined next by:

$$d_{ij,t} = \left\{\sum_{s=1}^{p}\left|\hat{z}_{si} - \hat{z}_{sj}\right|^t\right\}^{1/t} = \left\|\hat{z}_i - \hat{z}_{sj}\right\|_t, \quad (10)$$

Next, define $R_{i,qt}$ for a given set of clusters:

$$R_{i,qt}\left(U, \bar{V}, (U)\right) = \max_{j,j\neq i}\left\{\frac{S_{i,q}(U)+S_{i,q}(U)}{d_{ij,t}(U)}\right\}, \quad (11)$$

Finally the Davies-Bouldin index can be defined by:

$$V_{DB,qt}\left(U, \bar{V}, (U)\right) = \frac{1}{c}\sum_{i=1}^{c} R_{i,qt}\,(U), \quad (12)$$

Compact and well separated clusters are desirable, therefore, clustering occurs when the Davies-Bouldin index is small (Bezdek and Pal, 1998). Also, in defining clusters on the SOMs, the unified distance matrix (U-matrix) was used in conjunction with the Davies-Bouldin index. The U-matrix is a visual map of distances between neighboring map nodes and can help visually identify clusters (Vestano and Alhoniemi, 2000). For the factor clusters, pairwise Hotelling p-values between cluster means were compared to ensure that the newly formed clusters were significantly different from each other (Hammer, et al., 2009).

Cluster Interpretation Techniques

After clusters were defined, one tail t-tests were performed to compare water quality parameters in each cluster to the water quality of the entire watershed. These tests were run to determine if the mean values of water quality variables at a given cluster

were significantly larger or smaller than the mean values of water quality variables the entire dataset. T-tests assume that the parameters being tested have a normal distribution and equal variance (Davis, 2002). For this reason, the Box-Cox transformed variables were used in the comparison. Additionally, the Welch test statistic was used in cases where variance was unequal (Hammer et al., 2009). This is used over the traditional t-test because it does not employ a pooled variance estimate.

Classification Methodology

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is used in this research to predict the water quality cluster membership of any monitoring stations based on several quantitative spatial variables related to the stations's drainage basin (i.e. physical watershed characteristics). The main objectives of LDA are to determine a predictive equation that will classify an observation based on its set of spatial variables and to better understand the relationship between the discriminating variables and the clusters associated with them (similar to Santos-Roman et al., 2003). Stepwise discriminant analysis is a variable selection process used when there are several quantitative variables. This method is a useful precursor to direct parametric LDA (SAS, 2002-2004). Variables are chosen to enter the model according to the significance level of the F-test from an analysis of covariance (SAS, 2002-2004). The F-test gives an indication of how well a predictor variable discriminates between groups. Variables that exhibit the most discriminatory power are entered first, then the second most, and so on. This continues until all variables that meet a predetermined significance level are entered into the model. Additionally, variables are removed if their significance level drops below the predetermined criterion as more variables are entered into the model. For example, if the inclusion of variable A lowers the discriminating power of variable B below the significance level, variable B will be removed from the model. A moderate significance level of 0.10 to 0.25 is recommended by Costanza and Afifi (1979). When all variables still in the model meet the predetermined criterion, the stepwise selection is complete (SAS, 2002-2004). A classification equation is then determined by direct parametric LDA, which assigns stream/river monitoring sites into the determined water quality clusters. Classification equations are linear combinations of the predictor variables, and

these equations distinguish between different groups C_j of data (Tabachnick and Fidell, 1989). LDA classification equations take the following form:

$$C_j = c_{jo} + c_{j1}X_1 + c_{j2}X_2 + \ ... + c_{jp}X_p, \quad (13)$$

where X_p is a predictor variable, c_{jp} is a variable coefficient, and c_{jo} is a constant. Classification coefficients are determined with the means of the predictor variables **M** and the within-cluster variance covariance matrix,$\boldsymbol{W}$ (Tabachnick and Fidell, 1989).

$$\boldsymbol{C_j} = \boldsymbol{W'M_j}\ , \boldsymbol{C_j} = c_{j1}, c_{j2}, \dots\ , c_{jp} \quad (14)$$

These can be used as water quality prediction equations by inserting selected watershed characteristics into the equation (Santos-Roman, 2003). Stepwise LDA and direct parametric LDA were performed in SAS software.

Support Vector Machines

Support vector machine (SVM) classification is an alternative classification method to LDA. It is similar to another machine learning technique, the artificial neural network (ANN). They are both data-based modeling techniques, which learn relationships between input data (explanatory variables) and output data (response variables) with no previous knowledge of the underlying relationships between the data. The two modeling procedures even share the same architecture (Figure 4) (Kecman, 2001). The SVM has two attractive characteristics over the ANN. First, the SVM employs the structure risk minimization (SRM) principle, rather than the Empirical Risk Minimization (ERM). The SRM minimizes an upper bound on expected risk, rather than the error on the training data. This gives the SVM a greater ability to generalize, which is the ultimate goal in creating classification models (Gunn, 1998). Secondly, the training of the SVM is equivalent to training a linear model, but it can also identify non-linear patterns through the use of kernels (Ren et al., 2006). The kernels acts as a hidden layer that non-linearly maps input data into high dimensional space. The radial basis function kernel performs well with most types of data (Hsu et al., 2010).

Parameter selection is another key part of building an SVM. Essentially, a SVM classification tries to maximize the margin of a hyper-plane that is separating at least 2 groups of data. However, complete separation of the data can lead to poor generalization; therefore we employ the parameters γ and C. The γ parameter is a slack variable that allows the hyper-plane to not completely separate the parameters. The C parameter

decides the tradeoff between training error and the margin of the hyper-plane (Ren et al., 2006). The SVM models used in this work were obtained from LIBSVM – A Library for Support Vector Machines (Chang and Lin, 2000-2010) on a Matlab interface.

Cross Validation

Leave-one-out cross validation is used to test model performance especially when sample size is small. Often times, validation is performed by splitting a dataset into a training set and a testing set to derive the apparent error of a model (SAS, 2002-2004). For this study, many of the multivariate analyses require a large sample size, and dividing the limited number of monitoring stations into a training set and testing set was not practical. However, leave-one-out cross validation is an alternative way to test model performance, and it does not require a test set. In this study, leave-one-out cross validation was employed on both the LDA and SVM models to test their performance. Leave-one-out cross validation trains a SVM or LDA based on n-1 observations then, applies the model to the observation that was left out. It does this for all observations, and the misclassification rate indicates the performance of a given model (SAS, 2002-2004).

Cross validation served a dual purpose in training the SVM, as it indicated model performance and aided in the selection of the parameters C and γ. The Grid-search method as described by Hsu et al. (2010) was applied during SVM training. The Grid-search procedure is a straight forward procedure in which various combinations of C and γ are used in the SVM and the combination that produces the best cross validation is chosen. Sometimes different combinations of C and γ produced the same cross validation errors. These ties were broken by choosing the lower values of C, because it produces a better generalization of the model.

Wilcoxon Matched-Pairs Signed-Ranks Test

Once the classification models were built using the IDEM dataset, the next step was to test these models on the independently collected ECWMP dataset. The Wilcoxon matched-pairs signed-ranks test was used to determine whether the LDA or SVM was able to classify the unseen ECWMP data more accurately than the other. The Wilcoxon matched-pairs signed-ranks test is a non parametric test that is used to determine if a pair of data (e.g. LDA and SVM classification accuracy) is significantly different. SVM and

LDA classification accuracies on the unseen data were defined for each ECWMP station as the percentage of water quality variables whose values fell within the range of the cluster into which they were classified. The null hypothesis, H_0, for this test was chosen to be true if the accuracy of LDA and SVM were equivalent. To begin, this test first finds the differences between the LDA and SVM accuracies for matched-pairs of ECWMP stations. The absolute values of the differences are ranked from smallest to largest. Then a sign is assigned to the ranking based on if the difference was positive or negative. The absolute values of the ranks with the sign that appears the least are then summed. The sum of the ranks is the value T*, which is compared to a table of critical values of T. If T* is greater than the critical value of T at for a given sample size at a given significance level, then H_0 is rejected (Siegel, 1956).

RESULTS AND DISCUSSION

Variable Reduction

Principal Component Analysis

The principal component analysis (PCA) was performed on the time-averaged 16 water quality variables at the 44 IDEM water quality monitoring stations in the watershed. Before statistical analysis was conducted, the data was tested for normality with the Shapiro-Wilk goodness of fit test (Hammer et al., 2009). Variables that were not normal at an α=0.05 level were normalized using Box-Cox power transformations (Kutner et al., 2004). After the data was transformed it was checked for normality again. Box-Cox power transformation values and the results of the Shapiro-Wilk tests can be seen in Appendix A. Lastly, the data was scaled using the softmax transformation (SAS, 2002-2004). This last step was done to rescale the values of the 16 water quality variables in the datasets to similar scales and reduce the effect of any outliers that remained after normalization.

All 20 datasets were analyzed independently of one another. In 19 out of 20 datasets four factors (factors refer to the principal components) were retained from the PCA, and this was determined by examining the Kaiser criterion and Scree plot (Suhr, 2005). In all datasets, the first four factors explained 85% to 91% of the variance in the data. Each factor was examined for variables with the highest contribution or loading to the factor, and varimax rotation was used to better identify variables contributing to each factor (Suhr, 2005). Paul et al. (2006) selected variables with loadings over 0.6 to be associated with a given factor, in their work. Santos-Roman et al. (2003) considered factor loadings over 0.55, and Iscen et al. (2007) considered loadings over 0.5. For this study, variables with factor loadings >0.6 will be considered to have significant contribution to the associated factor. Variables that did not have factor loadings greater than 0.6 were removed from the PCA, since they could not be clearly associated with any of the factors (Suhr, 2005). Most of the variation between the 20 PCA results occurred between the annual and quarterly dataset, rather than between statistical indicators. However, statistical indicators did produce different PCA results, but these differences, generally, did not change the interpretation of the PCA.

The geometric mean is used as an example factor loading matrix for each of the annual and quarterly datasets. This was done to emphasize how seasonal changes affected variable reduction. Factor loadings matrices for the other statistical indicators can be located in Appendix A. The factor loadings for the annual geometric mean dataset are shown in Table 2. In the annual geometric mean dataset the first factor explained about 31% of the variance, and alkalinity, chloride, hardness, nitrate + nitrite, specific conductance, and sulfate had factor loadings greater than 0.6. These six variables distinguish themselves from the other variables because they are all transportable in groundwater, or in the case of alkalinity and specific conductance, are a measure of cations and anions that are concentrated in subsurface flow (Hem, 1985). Therefore, this factor is associated with subsurface flow. The second factor from the annual geometric mean data set had high loadings from total suspended solids, turbidity, iron, and temperature. Based on the first three variables, this factor can be associated with transport of suspended particles and their associated components (iron); it explains about 24% of the variance in the dataset. Though temperature loads high with this factor, it is difficult to explain any exclusive dependencies between particles and temperature, since multiple other causes such as, stream shading, the urban heat island, geospatial position, point source discharges, etc, can also affect temperature. This complexity can be seen in the seasonal datasets where temperature loadings behave erratically. The third factor has high loadings from total organic carbon (TOC), chemical oxygen demand (COD), and total Kjeldahl nitrogen (TKN) and accounts for about 21% of the variance in their respective datasets. TOC, COD, and TKN are closely related to the organics and organic pollutants in water (Hem, 1985). The fourth and final factor from the annual dataset explains about 11% of the variance and is associated with dissolved oxygen and pH. These variables describe the reduction/oxidation or redox conditions in the water, as well as the buffering capacity of water that is related to the underlying geology (Hem, 1985). While total phosphorus did not load highly any of the factors factor, the initial PCA runs showed that it had loadings greater 0.5 on the first 3 factors. This indicates that total phosphorus is a complex variable that cannot be associated simply with one factor and, therefore, had to be removed from the final PCA of the annual geometric mean dataset (Suhr, 2005).

The quarter 1 (January – March) datasets showed similar results to the annual dataset. Table 3 shows the factor loadings for the quarter 1 geometric mean dataset. Factor 1 had high loadings from the same subsurface flow-associated variables in the annual data set and explained about 33% of the variance in the data set. The second factor in quarter 1 was similar to the organic-associated factor in the annual dataset. However, it was always the second most important among statistical indicator datasets explaining 21% to 25% of the variance in each of the datasets. Also, total phosphorus loaded highly on this factor. The high loading of total phosphorus with the organic factor could be attributed to a winter and spring flushing phenomenon documented by Dalzell et al. (2006). In that study they examined TOC that builds up during the winter and is flushed out in high spring flows. Organic particulate phosphorus, one component of total phosphorus, exists in the plant material and manure that builds up over the winter months (Hem, 1985). Since quarter 1 covers January through March, a flushing effect from high flows in the late winter and early spring explains the high loading of total phosphorus with the organic factor in quarter 1. The third factor was similar to the particle-associated factor from the annual dataset. TSS, turbidity, and iron loaded highly on this factor for each statistical indicator. Temperature, however, was never associated with this factor during this time period (January – March). The particle-associated factor explained 19% to 21% of the variance among the datasets. The fourth factor for the quarter 1 geometric mean dataset was similar to the annual geometric mean dataset's fourth factor, as it was again associated with redox conditions in the water. Dissolved oxygen and pH loaded highly together in each instance. Temperature did not load highly on any factor for the quarter 1 geometric mean dataset, however, for the trimmed mean dataset (Appendix A), the inclusion of temperature added a caveat to the redox factor as it showed a high negative loading on this factor. This indicated that temperature has an opposite correlation with dissolved oxygen and pH. It also showed up as its own factor, explaining about 8% of the variance in the dataset, for the median dataset.

The quarter 2 (April – June) geometric mean dataset continued with the theme of the annual and quarter 1 datasets. The loadings for the quarter 2 geometric mean dataset can be seen in Tables 4. The first factor was the subsurface flow-associated factor, and explained about 31% of the variance in the dataset. The second factor for the quarter 2

geometric mean dataset was the organic-associated factor and explained about 25% of the variance in the dataset. The organic-associated factor was once again characterized by TOC, TKN, and COD, however total phosphorus did not load greater than 0.6 as it did in quarter 1. Rather, its loading behavior was complex, similar to the annual dataset. Again, it had fairly high loadings on the first 3 factors (but below the 0.6 criterion), and, due to this complexity, it was removed from the subsequent PCAs (Suhr, 2005). Temperature also had a high loading with the organic-associated factor in quarter 2. The particle-associated factor was the third most important factor, and it explained about 23% of the variance in the dataset. Once again this factor included TSS, turbidity, and iron. The redox condition-associated factor behaved the same as it did in the annual dataset, and it explained about 10% of the variance in the dataset.

The quarter 3 (July – September) and quarter 4 (October – November) PCA results were very similar. Their factor loadings can be seen in Table 5 and Table 6, respectively. Like the annual, quarter 1, and quarter 2 datasets, the first factor explained about 33% of the variance in the quarter 3 and quarter 4 datasets, and had high loadings from the same variables related to the subsurface flow-associated factor i.e. alkalinity, chloride, hardness, nitrate + nitrite, specific conductance, and sulfate. However, in both quarter 3 and quarter 4, total phosphorus loaded highly on this factor. As stated previously, phosphorus is most commonly transported with particulates and particle associated with organic matter, which does not fit well with the subsurface flow characterization of this factor (Hem, 1985). In the White River watershed dataset nearly all of the monitoring stations showed the highest total phosphorus concentrations in quarters 3 and 4. It is likely that the contribution of phosphorus to streams in the quarters 1 and 2, when flows are highest, is a result of the flushing effect of overland flow. Then in quarters 3 and 4, during low flow times, phosphorus concentrations increase due to a reduction in the dilution of point source phosphorus inputs, as well as in situ biological production. Since total phosphorus does not load highly on one factor in the quarter 2 dataset, it likely represents a time of event flows with increases in both particle-associated phosphorus, but also dilution from the increased precipitation and low-P groundwater inputs to streams. The second factor for the quarter 3 and 4 geometric mean datasets explained about 23% of the variance in the dataset. It was characterized as the

particle-associated factor, so TSS, turbidity, and iron all load highly on this factor. Water temperature also loads highly on this factor in the quarter 4 geometric mean dataset, but, because temperature is not related to particles in water, this factor is more accurately characterized as being related to particles plus temperature in quarter 4. Temperature does not meet the 0.6 criterion for any factor in the quarter 3 geometric mean dataset. The third factor explaining about 21% of the dataset for both the quarter 3 and 4 geometric mean datasets is the organic-related factor and includes, TOC, COD, and TKN. The fourth factor, explaining about 12% of the dataset, once again describes redox conditions in the water and includes high loading from dissolved oxygen and pH.

Kohonen Self-Organizing Map Results

Kohonen self-organizing maps (SOMs) were constructed for the annual and seasonal datasets for each statistical indicator. Before statistical analysis was conducted, each variable was scaled using the softmax transformation. This step ensured all variables were within the range [0, 1] and reduced the effect of outliers. The goal of the self-organizing map was to construct a two dimensional representation of the original 16 water quality variables. The maps were created using a hexagonal topology and a 13 node by 11 node architecture. The map's 143 nodes were given random initial values between 0 and 1. Then the learning algorithm was run sequentially by having each station's standardized water quality variables acting as an input vector. The main purpose of the learning algorithm was to organize the similar water quality stations using a technique known as vector quantization (Rojas, 1996). This process essentially projects the water quality variables from each monitoring station in 2-dimensional space. The algorithm was run for 5000 iterations. The first 1000 iterations were a rough training phase, which consisted of a neighborhood radius of 4 nodes and an initial learning rate of 0.5. The next 4000 iterations were the fine tuning phase, and the neighborhood radius for this phase was 1 node and the learning rate was initially 0.05 and was reduced to 0 as the learning finished.

Two useful visualizations of the SOM are the unified distance matrix (U-matrix) and the component maps for each individual variable. The component maps show where each individual variable has high and low values on the map. The individual components for the SOMs can be seen in Appendix B. In order to make sense of these maps, one may

expect to see similar patterns in the maps of associated variables, such as alkalinity and hardness. These maps are comparable to the PCA results since variables that loaded highly on a factor should have similar maps. Additionally, complex variables or variables that did not have high loadings from the PCA are included in the SOM. Ultimately, all of these component maps are combined to create the U-matrix, which characterizes the Euclidean distances between each node. The U-matrix can be very useful in visually clustering data, and will be discussed more in depth in the Cluster Analysis section (Vestano, 2000). Before analyzing these SOMs, it is important to note that each SOM must be looked at individually. Since node values on each map are generated randomly, each time a SOM is generated, the locations of stations on a map will change, but the relative distances between stations will stay the same.

In order to contrast with the geometric mean PCA loading tables, the SOM component maps for each of the statistical indicators in the annual datasets can be seen in Figures 5 to 8 (the quarterly SOM component maps are located in Appendix B). These maps display the differences in the influence that the choice of statistical indicator has on variable reduction. By examining Figures 5 to 8, one can observe that the choice of a robust (outliers have a minimal, in the case of the geometric mean, or no effect on the indicator, in the case of the median), non-robust (outliers potentially have a great effect on the indicator i.e. mean), or semi-robust (some outliers are removed but they can still have a great effect on the indicator i.e. trimmed mean) statistical indicator has an effect on variable reduction. For example, TSS, turbidity, and iron are all variables whose concentrations have outlying peaks during high flow conditions. On the other hand, while under low flow conditions, they have relatively small concentrations. By comparing the annual mean (non-robust statistical indicator) dataset SOM component map (Figure 5) and the annual geometric mean (robust statistical indicator) dataset SOM component map (Figure 8), one can observe the influence of these outlying values. For the mean SOM component maps relatively high values of TSS, turbidity, and total iron are spread out among a larger group of nodes. However, in the geometric mean SOM component maps the high values for these three variables are located in a much more compact group of nodes. The larger spread of high node values in mean dataset is likely a result of the large outlying values skewing the mean to a higher value than is actually

representative of a given monitoring station. On the other hand, those large outlying values are also included in calculation of the geometric mean, but their inclusion is not so obviously reflected in the geometric mean SOM component maps. Since the geometric mean rescales the data on a log scale while calculating the arithmetic average (i.e., $Geometric\ mean = Antilog\left[\frac{1}{n}\log x\right]$), the influence of large outlying values on the geometric mean is greatly reduced. The geometric mean is advantageous in that it is able to include all of the information from the dataset, without skewing its value towards the high outlying values that only occur during high flows. Also, because of the aspect of using all of the data for a given variable, the geometric mean was also considered to be superior to the median, which only uses the midpoint of the dataset. Lastly, the trimmed mean is considered a semi-robust indicator because it attempts to dampen the effect of large outliers by removing the highest and lowest 5% of the data. However, the decision of how much data to remove is arbitrary, and it is susceptible to removing too much data or not enough data depending on the scenario. The influence of robust, non-robust, and semi-robust statistical indicators can be observed in most of the water quality variables.

Aside from the effect of the choice of statistical indicator, the same annual and quarterly patterns seen in the PCA loading matrices were observed in the SOM component maps. However, the component maps also show that there is a more complex relationship between many of the variables than can be observed in the PCA loading matrices. Beginning with the variables that were characterized as subsurface flow-associated variables (i.e. alkalinity, chloride, hardness, nitrite + nitrate, specific conductance, and sulfate) alkalinity and hardness have nearly identical map patterns with the higher values on the same side of the map and lower values on the other. Nitrate + nitrite has a component map similar to those of alkalinity and hardness, but the highest values of nitrate + nitrite are located in a smaller area of the map. Chloride, specific conductance, and sulfate are also very similar to each other, but their general map pattern differs from alkalinity, hardness, and nitrate + nitrite. For example, in the annual mean dataset (Figure 5), the highest values of alkalinity, hardness, and nitrate + nitrite occur across the top half of the SOM. On the other hand, the highest values of chloride, specific conductance, and sulfate occur, in general, on the upper left hand side of the SOM. It is notable that the highest values of chloride and specific conductance share the

same cluster of nodes forming a "hot spot" on their respective component maps. The organic-associated factor variables (TOC, COD, and TKN) generally show the same pattern in all the annual SOMs. Although, there is a group of nodes where TOC and COD show higher values and TKN does not in the annual median, trimmed mean, and geometric mean datasets (Figures 6, 7,and 8). Interestingly, the component maps of chloride, sulfate, and specific conductance appear more similar to the component maps of the organic-associated variables than the component maps of alkalinity, hardness, and nitrate + nitrite. The particle-associated factor for the annual datasets (TSS, iron, turbidity, and temperature) also show the same general pattern between component maps, even though there is more variability in the temperature maps. The maps of the redox condition variables (pH and DO) showed one small area of low values in each of the annual component maps, while the rest of the maps were variable for pH and DO. Phosphorus is an interesting variable because it did not load highly on a variable in the PCA. Its component maps appear to be similar to both the organic associated variables and the chloride, specific conductance, and sulfate variables of the subsurface flow associated variables. This is a likely cause of the low loadings in the PCA, showing that phosphorus is a complex variable.

In the quarter 1 SOM, the subsurface flow-associated variables exhibited similar patterns to those that were observed in the annual dataset. Again, the organization of the SOMs for the quarter 1 datasets is similar to the factor pattern found in the PCA, except greater insight into the water quality variable behavior is achieved by examining the SOM component maps. While the PCA grouped water quality variables that loaded highly on a given factor, an examination of the component maps allows the viewer to more precisely observe how different water quality variables are behaving in relation to each other. Alkalinity and hardness have nearly identical component maps. Again, nitrate + nitrite is very similar, even though a fewer number of nodes are associated with the highest values of nitrate + nitrite. Chloride, sulfate, and specific conductance once again, have very similar component maps, with chloride and specific conductance sharing a "hot spot" of high values. Similar organization of the organic-associated variables was once again apparent. The component maps of TOC and COD were almost identical and TKN's component map followed a very similar pattern. The component map of

phosphorus is slightly different, but follows the same general pattern of TKN, TOC, and COD (phosphorus is usually part of the organic-associated factor in the quarter 1 PCA). The particle-associated variables are all mapped nearly identically. Temperature shares similar areas of high values as the particle associated variables, but its values are much more variable throughout the map. For the redox-associated variables, there is more of an apparent parallel between component maps in the quarter 1 SOMs than the annual SOMs. It is still difficult to discern, but a somewhat similar arrangement of high values can be observed between component maps. It is also apparent that the temperature component map has organized its high and low values opposite of the arrangement of DO and pH. This makes sense, because warmer waters will be able to hold less DO than colder waters. For a more in depth examination of the quarter 1 SOM, the component maps are located in Appendix B.

The quarter 2 SOMs fall into the same type of organization as the annual and quarter 1 SOMs. One notable change in the subsurface flow-associated variables is that the specific conductance component map has organized itself more closely to the map arrangements of alkalinity and hardness. Regarding the organic-associated variables, TOC and COD appear to have two areas of higher values on their respective maps, while TKN only shares one of those areas of higher values. The component maps for phosphorus, which did not exhibit a high loading on any factor in the PCA, is very similar to the maps for TKN and somewhat similar to the maps for chloride and sulfate. The particle-associated maps look nearly identical again. DO and pH are arranged in a somewhat ambiguous but similar manner. The temperature component map is arranged uniquely among all the component maps. Again the SOM component maps for quarter 2 are located in Appendix B.

The quarter 3 and quarter 4 SOM results are, in general, the same, but differ slightly from the other quarterly and annual datasets. Of note for the subsurface flow-associated variables is that the component maps for phosphorus is most similar to chloride, sulfate, and specific conductance. This visualization coordinates well with the findings of the PCA, where phosphorus loaded highly on this factor for all statistical indicators. The organic-associated factor variables remain consistent with the component maps for TOC and COD being the most similar and TKN sharing many of the

organization characteristics. The component maps for the particle-associated variables are once again nearly identical. The component maps for temperature are more variable, but, in general, they are organized in the same style of the other particle-associated variables. The component maps for DO and pH are probably more similar in the quarter 3 and quarter 4 datasets than in any other quarterly or annual dataset. Again the SOM component maps for quarters 3 and 4 are located in Appendix B.

Cluster Analysis

K-means cluster analyses were performed on the datasets in order to identify clusters of monitoring stations that had the most similar water quality. The cluster analyses were performed on the all of the datasets after variable reduction i.e. cluster analyses were performed on the retained factors from the PCA's and on the SOM's. The *K*-means method of clustering was used on both the factors and SOM's. This method of clustering is a partitional clustering method, which allowed for the factors or nodes to change cluster membership while the clustering algorithm ran. However, in partitional clustering, the initialization of the cluster centroids is random and can therefore create unlikely clusters (Rao and Srinivas, 2008). To minimize the chance of random clustering arrangements, 20000 iterations of the cluster analyses were run in order to find the best clustering arrangements.

Identifying Clusters

The next step in the cluster analysis was to decide the number of clusters that were present in the data. Besides acknowledging that physical watershed characteristics would affect the water quality of each station, no *a priori* knowledge of clusters was known going into the cluster analyses. Also, since the purpose of clustering the stations is to identify areas that have similar water quality conditions, the maximum amount of clusters was set to 10. If there were more than 10 clusters, it would create clusters with only one or two monitoring stations and defeat the purpose of creating generalized water quality clusters into which, unmonitored sites could be classified. In order to help determine the number of clusters for each analysis, the Davies-Bouldin index was examined. The Davies-Bouldin index helps indicate compact, well separated clusters, which is expressed by a ratio of within cluster scatter over between cluster separations (Bezdek, 1998). The index is calculated for each possible number of clusters, and the

best clustering configurations have a small index. Ideally, a plot of the Davies-Bouldin index would show a downward spike, followed by an increase in the index at the best clustering configuration. The plots of the Davies-Bouldin indices can be seen in Appendix B.

These downward spikes can be seen in several plots, but in many, there is a very dull spike or none at all. For these cases further investigation was needed to determine cluster membership. One example of an inconclusive Davies-Bouldin plot interpretation was for the quarter 1 geometric mean factor clusters where the Davies-Bouldin index fell sharply until 5 clusters, and then the slope of the plot leveled out (Figure 9). Since more than 5 clusters indicated a small improvement in cluster scatter over cluster separation, 5 clusters were chosen to generally represent the quarter 1 geometric mean factor dataset. Also, because of the ambiguity in cluster selection, pairwise Hotelling's tests were run to compare the different clusters created from the factors. The Hotelling's p-value indicated whether or not cluster means were different (Davis, 2002). Most clusters were distinct from each other at a significance level of $\alpha<0.05$. Results of the Hotelling's pairwise tests are in Appendix A. Not passing this test was likely a result of having too few stations in a given cluster. For example, the quarter 1 datasets generally had a large number of clusters. Because of this, some clusters had very few stations and did not pass the Hotelling's pairwise test. This was the only type of scenario where the clusters did not pass the Hotelling's test, therefore, it was determined that the cluster analysis, generally, created distinct clusters. The final cluster assignments for the clusters calculated by the factors in annual datasets are in Table 7. Additionally, the quarterly dataset factor cluster assignments are in Appendix A.

The Hotelling's pairwise test was not used as a confirmatory procedure for cluster presence for the clusters created from the SOM's. Rather, a visual examination of the unified distance matrix (U-matrix) combined with the *K*-means clustering method was used to determine the best clustering configuration. The U-Matrices generated by the SOM Toolbox 2.0 in Matlab are located in Appendix B (smaller versions of these U-matrices can be seen in the component maps as well). The nodes in the U-matrix are more than the nodes on the SOM's since they represent the distance between neighboring nodes on the SOM. Red nodes on the U-matrix represent a relatively farther distance

between nodes on the SOM, and the blue nodes represent shorter distances between nodes. Therefore, blue areas on the U-matrix represent tightly clustered nodes, and red areas represent separation of nodes on the SOM (Vestano, 2000). If only visual examination was used, one would identify clusters by identifying areas of blue surrounded by areas of red in the U-matrix. However, this procedure is tedious and inconsistent (Vestano, 2000). Therefore, the K-means procedure was used to cluster the data, then possible cluster configurations were compared with the corresponding U-matrix, and the final cluster configuration was chosen. For example, the Davies-Bouldin plot for the annual geometric mean SOM shows a downward peak at 6 clusters and 8 clusters. Both of these clustering arrangements were compared to the U- matrix to see if one corresponded better to the ideal scenario of tightly clustered blue nodes surrounded by red nodes (Figure 9). It was determined that the 8 cluster arrangement more closely follows the guidelines for clustering based on the U-matrix. In the 6 cluster scenario, the large yellow cluster appears to have three different sub-clusters according to the U-matrix. However, when the SOM is partitioned into 8 clusters, that large yellow cluster is broken up into those three sub-clusters, and the SOM more closely follow the U-matrix. The final cluster the numerical assignments annual dataset SOMs are in Table 8. The numerical cluster assignment tables for the quarterly SOM clusters are located in Appendix A. Also, the cluster configuration figures for each of the SOMs are located in Appendix B.

Interpreting the Clusters

With cluster membership determined, the next step in the analysis was to characterize each cluster based on water quality. To do this the mean of each water quality parameter in every cluster was compared to the mean value of that parameter among all the stations for a given dataset. For example, the mean value of alkalinity in cluster 1 from the annual geometric mean SOM cluster analysis was compared to the mean value of alkalinity among all the stations in the annual geometric mean dataset. Simple univariate one sided t tests were used to compare the means (Davis, 2002). Following the assumption of normality for the t tests, the Box-Cox transformed water quality datasets were used in the comparison. Additionally, the F-test for equal variances among the distribution of the cluster and the overall dataset was also reviewed. If this

test indicated unequal variances at an α=0.05, the Welch test for unequal variances was used instead (Hammer et al., 2009). If the t test indicated that a given cluster mean was higher or lower than the mean of the entire dataset at an α=0.05 significance level, that cluster was marked high or low, respectively. Additionally, cluster means that did not show a significant difference from the mean were marked moderate/variable (M/V). It was assumed that either these clusters did not differ much from the mean, or the parameter values in the cluster varied so much that it could not be stated with the given level of significance that the cluster mean was higher or lower than the mean of the entire dataset. Since there were 40 different cluster analyses, the annual geometric mean SOM clusters was used as an example to display the results from the cluster comparison t-tests and can be seen in Table 9. Appendix A contains all of the cluster t-test comparison tables.

In addition to the t tests, a more qualitative and visual interpretation of the clusters can be made by examining the component maps of the SOM's, and by looking at factor values of each cluster. By comparing the cluster configurations of a given SOM to the component maps of the corresponding SOM, one can make a qualitative and sometimes more descriptive explanation of a given cluster. For example, in cluster 4 of the annual geometric mean SOM, the t-test indicates that this cluster has a higher than average mean for sulfate when compared to the rest of the data set (Table 9). Figure 11 shows the component map of the annual geometric mean SOM next to its corresponding cluster arrangement. By examining Figure 11, one can see that the highest values (red/orange nodes) of cluster 4 are in only some stations (IWC-9, WR-248, WR-279, WR-293, WR-309), while the other stations in cluster 4 (EC-21, WLC-2, CIC-17, BL-64) have relatively moderate sulfate values (blue/green nodes).

The clusters created from the retained factors can also give further insight into cluster descriptions. When the factors are created in the PCA, they are standardized to have a mean of 0 and a standard deviation of 1 (SAS, 2002-2004). In order to visualize the clusters, the factors were divided into their respective clusters, and their respective means and standard deviations were plotted. Since factors are not the reduced versions of the original variables, these graphs will not be as descriptive as the component maps. However, these graphs can give a quick qualitative description of a cluster. For example,

by examining the plot for cluster 1 in the annual geometric mean factor clustering results, one can interpret a general assessment of cluster 1 (Figure12). From this box-plot, cluster 1 would be expected to have moderate to variable concentrations of the subsurface flow related variables, high concentrations of the particle-associated variables, slightly high concentrations of the organic related variables, and high but variable values for the redox-associated variables. In this example, one was able to observe that both the organic and particle-associated variables had high concentrations. However, it was evident that the particle-associated factor differed more from the mean of the entire dataset than the organic-associated factor did. The corresponding t-test table indicates that the particle and organic associated variables were high, but it does not give an assessment of the degree to which the organic and particle associated variables differ from the mean. The box-plots of each clusters' factors give the examiner a better idea how a certain factor is differing from the mean of the entire dataset. Plots of the factor means for each of the clusters from every factor cluster analysis can be seen in Appendix B.

While describing each cluster is important, the fact that there are so many clusters can make it difficult to interpret what is actually going on in the watershed. The variability in clustering is likely a result of a combination of factors. The selection of the number of clusters chosen, the variable reduction method, the time-averaging techniques, and the season represented by a given dataset all add to the variability of the clusters. The latter three are the most interesting for this research, since the selection of the number of clusters is based on the interpretation Davies-Bouldin index and somewhat arbitrary. Therefore, each cluster analysis result was compared and contrasted to try to identify patterns between quarterly datasets, statistical indicators, and variable reduction methods. Appendix A contains tables that show stations that always clustered together among different statistical indicators, among different quarters, and among variable reduction methods. By examining clustering patterns, stations that are sensitive to changes in statistical indicator, seasonal changes, and variable reduction method can be identified.

The fact that the clustering configurations change shows that changes in season, statistical indicators, and linear or nonlinear variable reduction are critical in finding similarity between water quality conditions at monitoring stations. Santos-Roman et al.

(2003) showed that using mean and median concentrations of water quality variables did not affect clustering. However, it is apparent that with the White River dataset robust statistical indicators versus non-robust statistical indicators plays a role in clustering. For example, stations GC-8, SND-4, and VF-38 always cluster together among different statistical indicators for the annual SOM clustering configurations. Additionally, IN-2, LST-2, MU-20, and SLT-12 always cluster together for the same datasets. However, for the non-robust mean and semi-robust trimmed mean datasets, all of these stations are put into the same clusters. It is only for median and geometric mean datasets, or the robust datasets that these stations are different enough to be put in separate clusters. Figure 13 shows this station-cluster shifting effect. In this instance, the calculation of the statistical indicator has likely affected cluster membership. When the mean and trimmed mean are calculated large outlying values are often included in the calculation, which can greatly skew the output. The trimmed mean attempts to remove the largest outliers, but selection of how much to remove is arbitrary and large outlying values can still be included in the calculation. On the other hand, when the median and geometric mean statistical indicators are calculated large outlying water quality values have a lesser impact (or no impact for the median) on the end result.

Dividing the datasets into different quarters also proved to be significant in cluster configuration. For example, when looking among the different quarterly and annual datasets for the SOM geometric mean clustering configurations, one can see that stations CIC-17 and EW-239 are in the same cluster for quarters 1 and 2 (clusters 1 and 3, respectively) and are in different clusters in quarters 3 and 4 (clusters 3 and 1, respectively for CIC-17 and clusters 7 and 5, respectively for EW-239). These station-cluster shifts can be seen in Figure 14. Both CIC-17 and EW-239 are characterized by high levels alkalinity and hardness and low to moderate concentrations of the particle-associated variables in all seasons. The differences between these stations in quarters 3 and 4 are primarily between concentrations of phosphorus and sulfate. In quarters 1 and 3, both stations show low to moderate concentrations of phosphorus and sulfate. These changes occur in quarters 3 and 4 when CIC-17 shows a relatively high concentration of sulfate and phosphorus while EW-239 maintains the low or moderate relative concentration of these variables. It was noted earlier that seasonal changes affect the

phosphorus transport pathways and could be a cause of these quarterly clustering shifts. Several more examples like this can be observed throughout the clustering configurations.

The clustering of the factors and the SOM proved to be similar but not the same. Many of the stations that consistently cluster together among different datasets are the same in the factor clusters as they are in the SOM clusters. However, the SOM clusters had more consistently clustered stations. Figure 15 shows a side by side comparison of the annual geometric mean clusters for the factor and SOM datasets. Across the different time-averaging statistical indicators, an average of 34 stations clustered with at least one other station each time for the factor clustering arrangements. On the other hand, an average of 38 stations clustered with at least one other station each time for the SOM clustering arrangements among different statistical indicators. For the different quarterly and annual datasets the factor clusters averaged 26 stations that clustered with at least one other station each time, while the SOM clusters averaged 33 stations. This indicates that the SOM was better able to detect similarities between similar stations, irrespective of the choice of the statistical indicators for time-averaging of water quality variables, or irrespective of the choice how the water quality dataset was reorganized into annual or seasonal datasets.

In general, clustered stations tend to vary more among different quarters than different statistical indicators. This is evident the fact that an average of 36 of the 44 stations clustered with at least one other station when cluster consistency varied among a single quarterly or annual dataset. On the other hand, by examining station cluster changes among the annual and quarterly datasets when the statistical indicator was held constant, it was determined that an average of 29 out of the 44 stations clustered with at least one other station. This result shows that an average of 7 more stations are sensitive to seasonal changes than changes in the choice of statistical indicator. SOM clustering configurations also proved to be more consistent than the factor clustering configurations. 17 stations clustered with at least one other station for all 20 datasets for the clustering of the SOM, while only 4 stations clustered with at least one other station each time for all 20 of the factor cluster configurations. Only stations WR-19 and WR-210 clustered together for all 40 of the clustering configurations. Lastly, while the tables describing

cluster consistency in Appendix A indicate the cluster analyses' sensitivity to changes in data preparation and data reduction, each individual station should be examined in order to understand the underlying causes of cluster membership shifts.

Spatial Distribution of the Clusters

Each of the cluster configurations can be seen in Appendix B. These maps combined with maps of the physical watershed variables can be used as an instrument to form ideas about the classification of these clusters. A visual inspection of the spatial variables can offer insight into how the different cluster configurations will be classified by the linear discriminant analysis (LDA) and support vector machines (SVMs). In the aforementioned example, CIC-17 and EW-239 were identified as similar stations in quarters 1 and as dissimilar stations in quarters 3. The spatial variables (except rainfall) do not change quarterly (e.g. the types of bedrock that underlay a watershed will stay the same year round); however, the influence of different spatial variables on water quality does change in different quarters.

Classification

Once clusters were defined, the next step in this study was to create classification models to predict cluster membership. Three steps were involved in classification: define physical watershed parameters, create linear and nonlinear models, and test the performance of these models with unseen data.

Spatial Data

Classification models are formed by using physical watershed attributes to predict cluster membership based on the clusters formed from the *K*-means clustering of the SOMs and PCA factors. In total, 38 physical watershed variables were considered, to discriminate between cluster memberships. These 38 variables can be broken down into 9 different categories: hydrologic/geomorphologic variables, climatic variables, Ecoregions, Natural Regions, bedrock geology, point sources, land use, land use change, and soil drainage (Table 10). These variables were chosen because they have been shown to influence water quality and the selected spatial data is readily available and easily calculated in ArcGIS. This will be useful in future uses of these models.

Many of the hydrologic/geomorphologic variables were derived directly with the ArcHydro tool. Some of these variables are self explanatory such as the longest flow

path and drainage area. The sum of streams is calculated as the combined length of all of the streams (as defined by the National Hydrologic Dataset (NHD)) in a given watershed. Network density was the sum of streams divided by the drainage area in a given watershed. Figure 16 helps visualize the aforementioned variables as it shows the actual NHD. The number of streams included in this dataset is far more extensive than what was needed to delineate the monitoring station watersheds. This figure also shows the increase in network density in the southern half of the watershed. The average slope percentage of a given watershed was included in the classification as well (Figure 17). The areas of higher slope appear to be very close to the areas of higher network density.

Temperature and precipitation were the only two climatic variables included in classification. The temperature map used in this study was developed through a partnership of the Natural Resources Conservation Service (NRCS), the National Water and Climate Center (NWCC), and the developers of PRISM (Parameter-elevation Regressions on Independent Slopes Model) at Oregon State University. The temperature map contains the mean annual temperature for the period from 1971-2000 (Figure 18). The precipitation dataset contained monthly precipitation method for the same time period and was developed by the U.S. Department of Agriculture. Since the precipitation data was monthly, raster math in ArcGIS toolbox was used to create an annual precipitation dataset and a precipitation dataset for each quarter. The annual precipitation map can be seen in Figure 19, while the quarterly precipitation maps are located in Appendix B. In general, precipitation and temperature values increase as one travels south in the watershed.

The EPA defined level III Ecoregions were another physical watershed parameter considered as a percentage of a given station's watershed area. Two level III Ecoregions make up most of the White River watershed: the Eastern Corn Belt and the Interior Plateau (Figure 20). The Interior Valleys and Hills Ecoregion also makes up a significant portion of the watershed. However, only a few stations had a percentage of this Ecoregion in their watersheds. For this reason it was considered an outlier and not included in the classification models. These regions were designed because the similarities of the ecosystems in these regions provide a framework for management, research, and assessment of nonpoint source pollution in a given region (Woods et al.,

1998). The Eastern Corn Belt Ecoregion is primarily a rolling till plain that has extensive corn, soybean, and livestock production, which has affected stream chemistry and turbidity. The Interior Plateau Ecoregion is a much more rugged terrain than the Eastern Corn Belt. Its soils developed from the underlying sandstone, siltstone, shale, and limestone, rather than the underlying till of the Eastern Corn Belt. The Interior Plateau has a mix of agricultural and forested land use (Woods et al., 1998). Karst topography is prevalent in some areas of the plateau, affecting ground water inputs to the streams.

An alternative regional ecosystem designation was designed by the Indiana Natural Heritage Data Center. Four natural regions make up most of the White River watershed: the Central Till Plain, Bluegrass, Highland Rim, and Shawnee Hills (Figure 21). The Southwestern Lowlands and Southern Bottomlands Natural Regions were also present in some of the monitoring stations' watersheds, but, like the Interior Valleys and Hills Ecoregion, these data points were considered outliers and not included in the analysis. The Central Till Plain and Bluegrass natural regions make up most of the Eastern Corn Belt Ecoregion and correspond well with two of the Eastern Corn Belt's sub-ecoregions. The Central Till Plain Natural Region roughly follows the outline of the Loamy High Lime Till Plains sub-ecoregion. The Central Till Plain Natural Region is primarily crop land underlain by a shallow ground water area (Fenelon, 1998). The Bluegrass Natural Region roughly follows the Pre-Wisconsinan Drift Plains sub-ecoregion, but it is not underlain by a shallow ground water area. The Highland Rim and Shawnee Hills Natural Regions roughly follow the outline of the Interior Plateau Ecoregion. However, they do not correspond with the sub-ecoregions of the Interior Plateau despite being based on ecosystem characteristics. Special interest will be taken in interpretation of the stepwise linear discriminant analysis (LDA), where the most discriminatory variables will be identified and put into the model. If either of these regional designations is more important in distinguishing water quality characteristics, it may be apparent in the variables chosen by the stepwise LDA.

In addition to ecosystem based settings examined, the geological settings of each watershed were also assessed in the classification of the water quality clusters. Six basic sedimentary geologic bedrock types underlay the White River watershed and were calculated as a percentage of a given station's watersheds drainage area: gray shale, a

mix of limestone and dolomite, limestone, a mix of sandstone and shale, siltstone, and a mix of sandstone, limestone, and shale (Figure 22). Geology will affect many stream parameters such as sediment load and dissolved solids concentrations. Streams running through areas of clastic sedimentary rocks i.e. sandstone, siltstone, and shale would be expected to have higher concentrations of suspended sediment related parameters such as, TSS. Groundwater influences will not be as prevalent here (Fetter, 2001). Streams running through limestone and dolomite would be expected to show higher concentrations of dissolved solids such as carbonate and magnesium. Groundwater influences from the Karst topography in the Interior Plateau Ecoregion, and the more homogeneous aquifers of the Central Till Plain Natural region will both indicate higher concentrations of parameters like alkalinity and hardness (Fetter, 2001).

The next set of spatial variables that was included in the classification models were different point source variables. The point sources considered were the number combined sewer overflows per square mile (CSO/mi^2), the number of confined animal feeding operations per square mile ($CAFO/mi^2$), and the sum of the allowed discharge at sites in the National Pollution Discharge Elimination System ($NPDES/mi^2$) (Figure 23). NPDES permits are given to any facility that discharges pollutants into a body of water. Pollutants can come from municipal and non-municipal sources (industrial and commercial facilities) and consist of toxic pollutants such as metals and manmade organic compounds to parameters such as phosphorus or total suspended solids (USEPA, 1996). The spatial parameter NPDES sum of permitted flow does not define the type of pollutant, so it will be difficult to relate that number to specific water quality parameters in a given cluster. However, CSOs and CAFOs, which are also types of NPDES facilities, will be strongly related to the organic parameters in this study i.e. TKN, COD, TOC, and total phosphorus (USEPA, 1996). They are also point sources of particular concern to the White River watershed. For these reasons, these point sources specifically are included in the classification models.

Another class of physical watershed variables was the type of land use as a percentage of a given station's watershed. Six types of land use were considered in this study: urban, cultivated crops, forest, pasture/grassland/scrubland, wetlands, and water (Figure 24). As evident by this map, the three most prevalent land use types are

cultivated crops, forest, and urban, making up 54.6%, 22.8%, and 10.2% of the total White River watershed, respectively. Pasture/grassland/scrubland, water, and wetlands respectively make up 9.3%, 1.8%, and 1.2% of the total watershed. These variables were derived from the National Land Cover 2001 Dataset (Homer et al., 2004). Land use becomes an important factor when interpreting non-point source pollution. The agricultural and urban areas are greatly affected by anthropogenic sources of pollution. For example, areas with predominant cultivated crop land use will have water quality pollution associated with nutrients applied to crops as fertilizers and sediment associated with fallow fields during winter (Fenelon, 1998). Urban areas are impacted by a mixture of both point and non-point source inputs. Point sources include sewers, waste water treatment plants, industrial waste sites, and landfills, and these sites are sources of organic compounds, trace elements, and nutrients (Fenelon, 1998). Forested areas and wetlands should be absent of anthropogenic pollution and act more as filters to water pollution. In addition to 2001 land use, land use change between 1992 and 2001 was also calculated as a percentage of a given station's watershed. This dataset was retrofit to provide more accurate land cover change data, since methods of data collection changed between 1992 and 2001 (Fry et al., 2009). The land use change variables include changes from agriculture to urban, agriculture to forest, urban to agriculture, urban to forest, forest to agriculture, and forest to urban.

The last set of physical watershed variables considered was soil drainage characteristics as a percentage of a given station's watershed. The soil drainage classes are: well to excessively drained, moderately well drained, somewhat poorly drained, and very poorly drained (Figure 25). This map reflects natural drainage conditions and was created by the U.S. Department of Agriculture (USDA, 2004). Soil drainage is related to the coarseness of a soil and the slope of the terrain. However, over half of the White River watershed's soils are modified by tile drains that artificially drain shallow groundwater areas (Fenelon, 1998). Typically only well drained soils would allow for shallow subsurface flow of parameters such as nitrate. However, the presence of tile drains in poorly drained soils will also allow for subsurface flow of these parameters through the tile drains. For this reason, the soil drainage characteristics variable will likely be less effective at cluster discrimination in areas that undergo artificial drainage.

However the importance of well drained soils to subsurface flow in areas without artificial drainage is sufficient to include it in the classification models.

Linear Discriminant Analysis

Linear discriminant analysis was the first method used to create classification equations based on spatial variables and cluster membership as defined by the clustering of the PCA factors. Stepwise linear discriminant analysis (LDA) was used as a prelude to creating the linear discriminant equations. This step helped identify the most discriminatory variables and reduces the problems caused by multicollinearity. The LDA assumes multivariate normality, but violations of this assumption are not fatal as long as non-normality is not caused by outliers (Tabachnick and Fidell, 1989). For this reason, all spatial variables were initially standardized using a logistic softmax transformation to reduce the effect of outliers. Variables were examined for the absence of outliers before they were inserted into the stepwise LDA. Stepwise selection was used at a 0.10 significance level. It must be noted that stepwise selection is not perfect because it selects variables solely on statistical criteria, rather than theoretical criteria (Tabachnick and Fidell, 1989). Therefore, caution must be taken in the interpretation of the results. The variables selected in the stepwise LDA are in Table 11.

Certain variables were selected more often than others among all of the spatial variables. According to the stepwise LDA, the following variables were significant in at least half of the analyses: Interior Plateau Ecoregion, NPDES, Highland Rim Natural Region, Shawnee Hills Natural Region, Water, Cultivated Crops, and Forest to Urban land use change. Of these, 1 is an Ecoregion, 1 is a point source, 2 are Natural Regions, 2 are land use variables, and 1 is a land use change variable. The stepwise LDA indicates that among the Ecoregions the areas off of the Eastern Corn Belt Ecoregion are most discriminatory i.e. the Interior Plateau. Further, the Highland Rim and Shawnee Hills Natural Regions, which make up parts of the Interior Plateau, are also very discriminatory. Although linking the NPDES variable to specific water quality variables is impossible, it is significant in discriminating between water quality clusters. Of the land use variables, water and cultivated crops, the cultivated crops presence makes sense in that it will add to pollutants to a stream from herbicide and pesticide use (Fenelon, 1998). The inclusion of the percentage water variable, while technically a land use

variable, could indicate stream density or the presence of lakes and reservoirs in a given watershed. Since there are several reservoirs in the watershed, this variable could also indicate the influence of reservoirs acting as sinks for several of the water quality variables. The land use change variable of forest to urban shows that the development of previously forested land has affected water quality, and can help distinguish cluster membership.

While some variables showed the most significance in distinguishing cluster membership, many variables were oftentimes removed during the stepwise LDA. The following variables were significant according to the stepwise LDA in less than a quarter of the datasets: Network Density, Slope Percentage, Precipitation, Eastern Corn Belt Natural Region, CSO/mi^2, Impervious Surface, Gray Shale, Limestone, Limestone/Dolomite, Siltstone, Bluegrass Natural Region, Urban to Forest land use change, Urban to Agriculture land use change, Forest to Agriculture, Well to Excessively Drained Soil, and Somewhat Poorly Drained Soil. These variables were likely inconsequential to water quality or redundant. For example, land use change from urban to another land use has not been nearly as prevalent, nor as consequential in the White River watershed as changes from agriculture to urban or forest to urban land use, which were significant in 9 and 12 of the 20 stepwise LDA's performed. Interestingly, the ecoregions and natural regions in the north eastern section of the watershed i.e. Eastern Corn Belt Ecoregion, Central Till Plain Natural Region, and the Bluegrass Natural Region proved to be less significant at distinguishing water quality clusters than their counterparts in the south central area of the watershed i.e. Interior Plateau Ecoregion, Highland Rim Natural Region, and Shawnee Hills Natural Region. These variables are inversely correlated, and therefore, inclusion of all of these variables would be redundant. Overall, the bedrock geology was inconsequential at distinguishing water quality clusters. It is likely that the changes in bedrock geology and the resulting effects on water quality were picked up by the ecoregions and/or the natural regions instead. Additionally, likely due to artificial drainage, natural soil drainage characteristics were not significant in many of the analyses. The CSO variable was likely redundant with the NPDES taking most of the credit in discrimination of the clusters. Of the climatic variables, precipitation was mostly inconsequential while temperature, which follows a similar

spatial pattern to the precipitation, was significant in 9 out of 20 of the analyses. The hydrologic variables, normally a function of stream size or terrain, showed up a moderate amount of times in the analyses, with drainage area the most common with nine appearances and slope percentage the least common with only 1 appearance in all of the stepwise LDAs.

Variable selection also showed some interesting patterns between quarterly datasets. The cluster analysis indicated that several stations were affected by quarterly changes. By examining variable selection in the quarterly dataset, the driving processes of these changes may become evident. For example, the forest land use variable only shows up in 8 of the 20 stepwise LDAs. However, it was significant in all 4 of the quarter 1 datasets' analyses. This indicates that forest land use plays a larger role in determining water quality in quarter 1 than in other quarters. The most common variables (significant in at least 3 out of 4 stepwise LDAs) in quarter 1 were: Drainage Area, Temperature, Interior Plateau Ecoregion, NPDES, Gray Shale, Shawnee Hills Natural Region, Forest, and Grassland/Pasture/Scrubland. In quarter 2 these variables were: Interior Plateau Ecoregion, NPDES, Highland Rim Natural Region, Grassland/Pasture/Scrubland, and Forest to Urban land use change. In quarter 3, these variables were: Longest Flow Path, Temperature, and Water. In quarter 4 these variables were Drainage Area, Interior Plateau Ecoregion, NPDES, High Rim Natural Region, and Shawnee Hills Natural Region. In all but quarter 3, there was at least one ecosystem based region among the most common spatial variables. Whether a Natural Region or Ecoregion, the ecosystem based regions proved to be a good separator of clusters. Additionally, variables related to stream size, i.e. drainage area and longest flow path, were most common in all but the quarter 2 analyses. Temperature proved to be a discriminant variable almost exclusively in quarters 1 and 3 – essentially winter and summer, respectively. The land use variables were generally unpredictable as to which one was more significant during the stepwise LDA. Only forest in quarter 1 and grassland/pasture/scrubland in quarters 1 and 2 were significant in at least 3 stepwise LDAs in each respective quarter.

Some of these results were also unexpected. For example, NPDES was a common variable in quarters 1, 2, and 4, but only showed up as significant in one of the

stepwise LDAs among quarter 3 data sets. The White River watershed generally follows a pattern of high flows in the winter and spring i.e. quarter 1 and quarter 2 and low flows in the summer and fall i.e. quarter 3 and quarter 4 (Fenelon, 1998). Point sources will have a much greater impact on stream water quality in low flow periods than in high flow periods, and non-point sources will be the main source of pollutants during high flow periods (Fenelon, 1998). This study mainly focuses on non-point sources, but does include 3 point source variables with the goal of capturing their effect on quarters 3 and 4 – the low flow period. However, the point source variables were more often significant in quarters 1 and 2 than they were in quarters 3 and 4. This reinforces the caution one must take in the interpretation of these statistically selected variables (Tabachnick and Fidell, 1989).

Once variable selection was complete, normal parametric linear discriminant analysis was performed on the selected spatial variables in an attempt to classify stations into their assigned cluster membership. Table 12 shows that problems with multicollinearity or singularity are unlikely since the pooled covariance rank is equal to the number of variables in every LDA (Tabachnick and Fidell, 1989). The LDA created classification equations for every possible cluster. These classification equations are analogous to multiple regressions, with cluster score acting as the dependent variable and the spatial variables acting as dependent variables. Each score is therefore a linear combination of the constant and each coefficient multiplied by its given spatial variable. For example, the classification equation for the annual geometric mean dataset can be seen in Table 13. The classification equation that results in the highest score indicates cluster membership for a given watershed based on its spatial variables. The classification equations for the all of LDA datasets are located in Appendix A.

In order to test the accuracy of these models leave-one-out cross validation was used. Another method of testing accuracy of the model would be to split the data into a training set and a testing set. However, with a small sample size (n=44), it was determined that cross validation would provide a more accurate reflection of model accuracy. Table 14 shows the percentage of stations correctly classified after cross validation. Among all models, the quarter 2 geometric mean and quarter 2 trimmed mean models have the highest accuracy, by correctly classifying 40 of the 44 stations (90.9%).

The quarter 4 trimmed mean model performed the worst, by correctly classifying only 27 of the 44 stations (61.4%). The average cross validation accuracy was 77.4%, or, on average, about 34 of the 44 monitoring stations were correctly classified after cross validation.

Support Vector Machine Results

Support vector machines (SVMs) were used to accomplish the same goal of the LDA: classification of water quality clusters based on physical watershed attributes. However, the SVM differs from the LDA in that it can express non-linear relationships between the spatial parameters and the cluster classification scheme, whereas the LDA is simple a linear combination of discriminating spatial variables. Additionally, the SVM was used to form classification models based on the clusters formed from the non-linear SOMs. The performance of the support vector machine was based on the selection of three parameters: kernel type, a regularization parameter C, and a training constant γ.

Kernel selection was the first step in constructing the SVM. The radial basis function (RBF) kernel was chosen for several reasons. First, it can perform the same tasks as the linear kernel, but can also deal with non-linear class and feature relationships. Secondly, the RBF comes with less numerical difficulty and has fewer hyperparameters e.g. C and γ to deal with than the polynomial kernel (Hsu et al., 2010). In general, the RBF kernel selection is most appropriate for the given problem and expertise of the practitioners (Hsu et al., 2010).

The selection of hyperparameters C and γ was accomplished by performing a grid-search. This method optimizes the parameters C and γ by running the model using leave-one-out cross validation. The goal of choosing appropriate values of C and λ is to find a balance between building a model that is too general and a model that is over-fitted for the training data (Ren et al., 2006). Different combinations of C and γ, with values ranging from 0.1 to 1000 and 0.0001 to 10, respectively, were used in the model. Values of C and γ were chosen from the models that had the best cross-validation accuracies (Table 15) (Hsu et al., 2010).

Feature selection was the next step considered in building the SVM. Stepwise LDA was used in the linear analysis to produce a subset of predictor variables; however, there is currently not a standardized method for variable selection in SVMs. Chen and

Lin (2006) proposed several feature selection strategies. Initial trials using their proposed F-score + Random Forest feature selection strategy did improve model accuracy on this dataset. Kartoun et al. (2006) used an optimal feature selection strategy where all possible combinations of 9 different features were selected. This resulted in only marginal improvement in their cross validation accuracies, and this technique did not seem practical in this study's case where there are 38 different features. Nilsson et al. (2006) showed that, while feature selection techniques have improved classification at low dimensions with features << samples, there is no such improvement at high dimensions. Given the high dimensionality of this data set, it was determined that feature selection was an unnecessary step.

Table 14 summarizes the performance of the SVM models according to leave-one-out cross validation accuracy. The best performing SVM was the annual mean dataset, and it correctly classified 41 out of the 44 stations (93.2%) using leave-one-out cross validation. The worst performing model was the quarter 3 median data set, which only correctly classified 27 of the 44 stations during cross validation. The average cross validation accuracy among all the models was 78.9% for the SVM models.

Comparison of SVM and LDA

Overall, the SVM slightly outperformed the LDA with an average cross validation accuracy of 79.7% to 77.4%. However, for different quarterly datasets LDA occasionally outperformed the SVM. SVM outperformed LDA in the annual, quarter 1, quarter 3, and quarter 4 datasets with average cross validation accuracies of 84.6% to 76.7%, 79.5% to 74.4%, 78.4% to 77.8% and 77.8% to 69.9%, respectively. However, LDA outperformed the SVM in quarter 2 with an average cross validation accuracie of 88.1% to 77.8%. From these cross validation accuracies, it is difficult to discern whether the SVM or the LDA models have greater predictor power.

In addition to cross validation accuracies, the models' performance for resubstitution can be examined station by station to compare SVM and LDA models. Resubstitution differs from cross validation in that all of the stations are simply input back into the model. Therefore, it is expected that the model will perform well since it has seen all of this data already. However, by examining which stations are classified incorrectly among the different models allows the user to identify possible weaknesses in

the model. The most commonly misclassified stations for the LDA were: EC-21, WR-319, FC-26, CIC-17, WR-279, and EW-79. For the SVM the most commonly misclassified stations were: EW-168, EC-7, and FC-7. All of these stations were misclassified in at least 4 of the models for the LDA and SVM, respectively. Table 16 shows all of the misclassified stations, and the clusters from which they were misclassified. Initially patterns among spatial variables were examined to see if a particular type of station was likely to be misclassified, such as stations with relatively small drainage areas or high urban land use. However, no pattern like this was apparent. Most misclassifications occurred when clusters were spatially close together. Additionally, smaller clusters were often clustered into larger clusters. Unbalanced clusters can cause a bias towards classification into the larger clusters for LDA and SVM (Tabachnick and Fidell, 1989, Tang et al., 2002). For example, in the quarter 1 geometric mean SVM model misclassified EC-1, EC-7, FC-0.6, FC-7, and WR-319, all into cluster 7 from their respective clusters. By comparing Table 16 and the spatial distribution of the quarter 1 geometric mean SOM clusters in Figure 14, one can observe that all of cluster 6 i.e. EC-1, EC-7, and FC-0.6 and cluster 4 i.e. FC-7 and WR-319 have been classified into the spatially nearby and larger cluster 7. Despite, the models not being perfect, both the SVM and LDA appear to do a good job at classifying stations into the created water quality stations based on physical watershed parameters.

Testing Models

Once these classification models were developed, the final step in this process was to see how well these models performed on unseen data. However, in order to include as much information as possible in creating the models, all of the IDEM monitoring stations were used as training data and no stations from this dataset were left over for testing. This problem was solved by using water quality datasets of Eagle Creek Watershed collected and published by the Center for Earth and Environmental Science (CEES) at IUPUI. The Eagle Creek Watershed Alliance (ECWA) has been conducting monthly monitoring at 11 different stations in the Eagle Creek watershed from March 2007 – present (March 2010 was the last update to the dataset at the time of this study). Within the IDEM dataset, the Eagle Creek watershed is already represented by 3 stations: EC-1, EC-7, and EC-21, and one would expect the performance of these stations to be

similar to those already in this watershed. However, there are some key differences between the ECWA sites and the IDEM sites that must be acknowledged. First, the ECWA sampling focuses on a much smaller area, and therefore many of the watersheds are smaller than those in the IDEM dataset (Figure 26). Additionally, all of these stations are located directly upstream of the Eagle Creek Reservoir. Only ECWMP-03 and ECWMP-04 are larger than any of the watersheds in the IDEM dataset. These two watersheds are very similar to the EC-21 site, which is also upstream of the Eagle Creek Reservoir. Also, there is less historical water quality data in the ECWMP dataset, since sampling started in 2007 rather than 1991. This will have a great affect on the impact of land use change on water quality. Additionally, only 14 of the 16 original water quality variables from the IDEM dataset are included in the ECWMP dataset.

Consistency was the key in preparation of the physical watershed parameters. Watersheds for each ECWA monitoring station were delineated using the Arc Hydro tool. The same physical watershed parameters in Table 11 were described according to the defined watershed for each station. Special attention was paid to scaling these spatial parameters. Scaling of the original watershed parameters was done to reduce numerical difficulties in calculations and so that parameters in high numeric ranges did not dominate those in small numeric ranges (Hsu, 2010). For classification to work, the ECWA watershed parameters had to be in the same scale as the original dataset. In order to do this each ECWMP station was scaled with the 44 original IDEM one at a time using the logistic softmax transformation. Therefore, the scaling of variable was accomplished by applying equation (6) to the 44 IDEM water quality variables plus an additional set of ECWA water quality variables. By doing this one at a time, the new ECWA stations were not affected by values of other ECWA stations, and were ready to be put into the respective SVM and LDA models.

Although the ECWA stations are confined to a small area, there is some variability among physical watershed characteristics. Land use, soil drainage characteristics, point sources, and bedrock geology exhibited the most spatial variability among the watersheds. Figures 27 to 30 show the variability in the ECWA watersheds of land use, bedrock geology, soil drainage, and point sources, respectively. All of the ECWA watersheds lie in the Central Till Plain Natural Region and the Eastern Corn Belt

Ecoregion. Also, because of the small area there is little variation of climatic and hydrologic parameters.

ECWA site classification was performed by simply inserting the scaled spatial data for each watershed into each of the LDA and SVM models. The output gives the cluster membership predictions and posterior probability estimates for each ECWA site. Tables 17 and 18 show the classification results of the LDA and SVM annual dataset models, respectively (The results for all of the datasets are located in Appendix A). The posterior probability estimates for the LDA and SVM are computed based on cross validation (SAS, 2004; Chang and Lin, 2001). These probability estimates indicate the percentage of times a given ECWA station was classified into a given cluster during cross validation. The cluster with largest probability estimate was the cluster into which the ECWA station was classified. In order to evaluate the performance of the classification models, the ECWA water quality parameters were compared to the IDEM water quality parameters in the cluster each ECWA station was assigned to. A spatial comparison of the annual geometric mean LDA and SVM classification can be made in Figures 31 and 32, respectively. In general these two models classified the ECWA stations into clusters that are located in the northern half of the White River watershed, which is in accordance to the location of the Eagle Creek watershed. Another interesting occurrence between these two models was that the SVM model classified ECWA station ECWMP-03 into the same cluster that the IDEM station EC-21 belonged to. However, the LDA model classified ECWMP-03 into a different cluster than the cluster that contained EC-21. This is interesting because EC-21 and ECWMP-03 are nearly identical watersheds, and one would logically assume that they would cluster together. In addition to a spatial comparison, a quantitative comparison between water quality at each ECWA station and the water quality of the respective clusters into which they were classified was performed. To do this, the range of the IDEM water quality parameter values for each predefined cluster was considered. Then it was determined if the ECWA stations' water quality parameter values fell within the range of the cluster for which it was classified into. The percentage of variables within the cluster range for each model was then calculated. Table 19 shows the results for this test for the annual dataset models. Cluster accuracy results for all of the models can be found in Appendix A.

Lastly, the performance of the SVM and LDA models were compared based on their respective cluster range accuracies with the Wilcoxon matched-pairs signed-ranks test. The cluster range accuracy was a measure of how similar the water quality a given ECWA was to the water quality of the cluster into which the given station was assigned. For each model and at each ECWA station the percentage of ECWA water quality variables that fell within the minimum-maximum range of the IDEM water quality variables in the assigned cluster defined cluster range accuracy. The Wilcoxon matched-pairs signed-ranks test is a nonparametric test that compares the cluster range accuracies of the ECWA stations for corresponding SVM and LDA models e.g. Annual Mean SVM model and the Annual Mean LDA model. It ranks the magnitude of the difference, and indicates if one model is superior to another (Siegel, 1956). The level of significance for this test was $\alpha=0.05$. Table 20 shows the decision matrix created by the results of the Wilcoxon matched-pairs signed-ranks test. According to this test, the SVM model had better success with cluster range accuracy for the annual mean dataset, the quarter 1 mean dataset, the quarter 3 mean dataset, the quarter 1 trimmed mean dataset, the quarter 2 trimmed mean dataset, and the quarter 3 trimmed mean dataset. The LDA had better success with only the quarter 2 mean dataset and quarter 2 median dataset. Neither the SVM nor LDA models had an advantage for the remaining 12 datasets. The cluster range accuracy and Wilcoxon matched-pairs signed-ranks test gives a slight advantage to the SVM models over the LDA models.

CONCLUSION

This study investigated different methods to construct linear and non-linear empirical classification models. The same set water quality monitoring stations were represented by different datasets that reflected different time-averaging techniques (by using statistical indicators such as mean, median, etc.), and temporal changes in water quality at different time scales (e.g. annual and quarterly). For each of these datasets, water quality monitoring stations were clustered into groups. This was accomplished only after the dimensions of the original water quality variables were reduced using a linear variable reduction method, PCA, and a non-linear variable reduction method, SOM. The PCA identified the 4 most important factors representing water quality, and the SOM projected the water quality variables in 2-dimensional space. Based on the PCA, the water quality variables could be broken down into four groups: subsurface flow-associated variables, organic-associated variables, sediment-associated variables, and redox condition-associated variables.

Clustering based on the PCA factors and the SOM showed that both statistical indicator and the quarter of the year a water quality sample was taken affected cluster membership. However, the differences in clustering between quarterly datasets showed that temporal changes had more of an effect on cluster membership. Nutrient loading in the streams in different seasons was shown to be one of the drivers causing the cluster membership shifts. There was also a difference noticed in the clustering of the SOM and PCA factors. It was noted that the clusters created by the SOMs across statistical indicators and different quarters were less variable than those created by clustering the factors.

After clustering, LDA and SVM were then used to create empirical classification models based on physical watershed data and cluster membership. These models were applied to unseen data from the ECWA. The Wilcoxon matched-pairs signed-ranks test showed that the SVM models classified the ECWA stations into clusters that more accurately reflected their water quality conditions in 6 out of the 20 possible models when compared to the equivalent LDA model. Conversely, the LDA outperformed the SVM in 2 out of the 20 possible models. In 12 out of the 20 models neither the SVM nor LDA did a better job at classifying the ECWA stations.

The objectives of this study were to compare the models that were built based on (1) statistical indicator, (2) annual or quarterly data, and (3) a linear or non-linear model. The choice of statistical indicator did appear to influence cluster membership of several of the water quality monitoring stations. Additionally, in classification, the geometric mean based models had an average cross validation accuracy of 80.2% compared to an average of 77.7%, 77.5%, and 78.6% for the mean, trimmed mean, and median models, respectively. Although the geometric mean did not greatly outperform the other statistical indicators, it is likely the most effective technique for time averaging long term water quality data, because of its ability to include all of the information from a given dataset and reduce the influence of outliers. While dividing data into quarterly subsets shifted cluster membership for many stations, this did not necessarily improve classification. In fact the annual models had the second best average cross accuracy error when compared to the quarterly models. Lastly, SVM slightly outperformed the LDA according to the average cross validation accuracy among all the models. The average cross validation accuracies for the SVM and LDA were 79.7% and 77.4%, respectively. However, the SVM outperformed the LDA in the Wilcoxon matched-pairs signed rank test, as well. In general, this study achieved its best results when using a non-linear classification model based on water quality data that was time averaged using the geometric mean.

Limitations in this study largely resulted from data limitations. All of the data included in this study was collected without this specific study in mind, and therefore, it did not necessarily conform to the demands of this study. For example, the land use change parameters reflected land use change from 1992 – 2001, while the IDEM water quality data had been collected from 1991 – 2008. Additionally, the ECWA test set of data was collected from 2007 – 2010, therefore comparing these two datasets must be done with knowledge of this in mind. Ideally, the dataset would have had larger sample size (e.g. n=150) at sites randomly located throughout the watershed. This would have better met sample size recommendation for the PCA and CA and allowed for the dataset to be divided into training and testing for the LDA and SVM without sacrificing the model learning. Furthermore, research is ongoing in the field of machine learning and future developments in SVM techniques will likely lead to more accurate models.

TABLES

Table 1 – Water quality variables selected for analysis

Alkalinity as $CaCO_3$ (mg/L)	Nitrite + Nitrate (NO_2 + NO_3) (mg/L)
Total Organic Carbon (TOC) (mg/L)	pH (SU)
Chloride (mg/L)	Total Phosphorus (mg/L)
Chemical Oxygen Demand (COD) (mg/L)	Total Suspended Solids (TSS) (mg/L)
Dissolved Oxygen (DO) (mg/L)	Specific Conductance (SC) (µS/cm)
Hardness (Ca + Mg) (mg/L)	Sulfate (SO_4) (mg/L)
Total Iron (mg/L)	Water Temperature (K)
Total Kjeldahl Nitrogen (TKN) (mg/L)	Turbidity (NTU)

Table 2 – Annual Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**86**	-32	-25	14	
TOC	-6	3	**95**	-5	
Chloride	**82**	-4	45	20	
COD	-3	39	**90**	7	
DO	16	-23	26	**82**	
Hardness	**91**	-30	-17	11	
TKN	30	37	**84**	12	
NO2 + NO3	**83**	-10	-19	7	
pH	12	10	-12	**89**	
Total P	DNL	DNL	DNL	DNL	
TSS	-4	**97**	15	-2	
SC	**95**	-11	22	14	
Sulfate	**81**	19	35	-9	
Temperature	-18	**75**	40	16	
Turbidity	-16	**94**	11	-4	
Iron	-14	**94**	10	-21	Final Communality
Eigenvalue	4.712	3.878	3.208	1.663	13.46
% Variance Explained	31.4	25.9	21.4	11.1	89.73333333

Table 3 – Quarter 1 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**91**	-11	-31	11	
TOC	-10	**96**	6	-9	
Chloride	**79**	51	-11	-1	
COD	-3	**92**	32	-11	
DO	-11	-10	-25	**84**	
Hardness	**92**	-10	-30	14	
TKN	37	**84**	29	-12	
NO2 + NO3	**73**	-35	2	25	
pH	30	-10	-22	**71**	
Total P	42	**63**	57	-4	
TSS	-8	18	**90**	-29	
SC	**91**	27	-9	3	
Sulfate	**79**	40	5	-23	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-22	29	**87**	-18	
Iron	-32	8	**85**	-17	Final Communality
Eigenvalue	4.898	3.653	3.116	1.541	13.208
% Variance Explained	32.7	24.4	20.8	10.3	88.05333333

Table 4 – Quarter 2 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**88**	-25	-24	16	
TOC	-6	**89**	-3	-7	
Chloride	**80**	46	-24	8	
COD	1	**91**	34	1	
DO	34	26	-40	**63**	
Hardness	**92**	-22	-23	13	
TKN	33	**85**	36	6	
NO2 + NO3	**71**	-43	11	24	
pH	7	-3	6	**92**	
Total P	DNL	DNL	DNL	DNL	
TSS	-8	26	**93**	-3	
SC	**91**	12	-16	12	
Sulfate	**82**	36	4	-20	
Temperature	-20	**75**	37	22	
Turbidity	-23	16	**93**	5	
Iron	-12	19	**95**	-13	Final Communality
Eigenvalue	4.716	3.74	3.376	1.476	13.309
% Variance Explained	31.4	24.9	22.5	9.8	88.72666667

Table 5 – Quarter 3 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**81**	-26	-39	13	
TOC	1	-8	**95**	-5	
Chloride	**83**	0	38	27	
COD	-3	38	**90**	18	
DO	14	1	19	**90**	
Hardness	**90**	-25	-26	5	
TKN	23	37	**83**	28	
NO2 + NO3	**84**	-14	-12	-1	
pH	12	13	1	**91**	
Total P	**71**	30	48	11	
TSS	-2	**96**	20	15	
SC	**94**	-7	21	16	
Sulfate	**82**	21	34	2	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-11	**96**	10	8	
Iron	-11	**97**	11	-5	Final Communality
Eigenvalue	5.044	3.368	3.274	1.911	13.597
% Variance Explained	33.6	22.5	21.8	12.7	90.64666667

Table 6 – Quarter 4 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**78**	-40	-33	13	
TOC	-5	5	**96**	-13	
Chloride	**86**	2	36	10	
COD	-1	37	**91**	0	
DO	-2	-29	3	**88**	
Hardness	**87**	-31	-25	14	
TKN	35	40	**79**	4	
NO2 + NO3	**83**	-3	-12	7	
pH	17	10	-10	**89**	
Total P	**67**	40	44	-1	
TSS	2	**96**	13	1	
SC	**96**	-5	17	6	
Sulfate	**83**	23	24	-14	
Temperature	24	**73**	31	-5	
Turbidity	-18	**93**	18	-1	
Iron	-21	**88**	14	-23	Final Communality
Eigenvalue	5.136	3.973	3.138	1.707	13.955
% Variance Explained	32.1	24.8	19.6	10.7	87.21875

Table 7 – Annual Factor Cluster Assignments

Station name	Geomean (5)	Mean (6)	Median (5)	Trimmed Mean (5)
BL-.7	5	5	3	1
BL-64	2	2	1	3
BWC-4	5	5	3	5
CIC-17	2	5	1	1
EC-1	4	3	2	2
EC-21	2	2	1	4
EC-7	4	3	2	4
EEL-1	3	1	5	3
EEL-38	3	1	5	3
EW-1	3	1	5	1
EW-168	5	1	3	1
EW-239	5	5	3	1
EW-79	3	1	5	3
EW-94	3	1	5	1
FC-0.6	4	3	2	2
FC-26	5	5	3	4
FC-7	4	3	2	2
FR-17	5	5	3	1
FR-64	5	5	3	4
GC-8	4	3	2	5
IN-2	3	4	5	3
IWC-9	2	2	1	2
LST-2	3	4	5	5
MC-18	3	1	5	3
MC-35	5	5	3	4
MU-20	3	4	5	4
SGR-1	5	5	3	5
SLT-12	3	4	5	3
SND-4	3	1	5	5
VF-38	4	3	2	3
WLC-2	5	5	3	4
WR-134	1	6	4	5
WR-162	1	6	4	4
WR-19	1	6	4	3
WR-192	2	2	1	4
WR-210	2	2	1	4
WR-248	2	2	1	2
WR-279	2	6	3	5
WR-293	2	2	1	1
WR-309	2	2	1	1
WR-319	2	1	1	3
WR-348	5	5	3	3
WR-46	1	6	4	1
WR-81	1	6	4	3

Table 8 – Annual SOM Cluster Assignments

Station name	Geomean (8)	Mean (3)	Median (8)	Trimmed Mean (7)
BL-.7	5	1	2	4
BL-64	4	1	4	6
BWC-4	5	2	2	4
CIC-17	4	1	4	6
EC-1	3	1	6	3
EC-21	4	1	4	6
EC-7	3	1	6	3
EEL-1	1	2	7	2
EEL-38	1	2	7	2
EW-1	1	2	7	2
EW-168	5	2	2	4
EW-239	5	1	2	4
EW-79	1	2	7	2
EW-94	1	2	7	2
FC-0.6	3	1	6	3
FC-26	5	1	2	4
FC-7	3	1	6	3
FR-17	5	1	2	4
FR-64	5	1	2	4
GC-8	2	2	5	7
IN-2	8	2	3	7
IWC-9	4	1	4	6
LST-2	8	2	3	7
MC-18	5	2	2	4
MC-35	5	1	2	4
MU-20	8	2	3	7
SGR-1	5	1	2	4
SLT-12	8	2	3	7
SND-4	2	2	5	7
VF-38	2	2	5	7
WLC-2	4	1	4	6
WR-134	6	3	1	5
WR-162	6	3	8	1
WR-19	6	3	1	5
WR-192	7	3	8	1
WR-210	7	3	8	1
WR-248	4	1	4	6
WR-279	4	3	4	6
WR-293	4	1	4	6
WR-309	4	1	4	6
WR-319	5	1	2	4
WR-348	5	1	2	4
WR-46	6	3	1	5
WR-81	6	3	1	5

Table 9 – The Annual Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Geomean SOM	LOW	LOW	M/V	HIGH	HIGH	M/V	M/V	LOW		
Total Organic Carbon	Annual Geomean SOM	M/V	HIGH	M/V	M/V	LOW	HIGH	HIGH	M/V		
Chloride	Annual Geomean SOM	LOW	LOW	M/V	HIGH	M/V	M/V	HIGH	LOW		
Chemical Oxygen Demand	Annual Geomean SOM	M/V	M/V	M/V	M/V	LOW	HIGH	HIGH	M/V		
Dissolved Oxygen	Annual Geomean SOM	M/V	M/V	HIGH	M/V	M/V	M/V	M/V	LOW		
Hardness	Annual Geomean SOM	LOW	LOW	M/V	HIGH	HIGH	M/V	M/V	LOW		
Total Kjeldahl Nitrogen	Annual Geomean SOM	M/V	M/V	M/V	M/V	LOW	HIGH	HIGH	LOW		
Nitrate + Nitrite	Annual Geomean SOM	LOW	M/V	M/V	HIGH	M/V	M/V	M/V	LOW		
pH	Annual Geomean SOM	M/V	M/V	LOW	M/V	M/V	HIGH	M/V	LOW		
Total Phosphorus	Annual Geomean SOM	M/V	M/V	M/V	HIGH	LOW	HIGH	HIGH	M/V		
Total Suspended Solids	Annual Geomean SOM	HIGH	M/V	M/V	M/V	M/V	HIGH	M/V	M/V		
Specific Conductance	Annual Geomean SOM	LOW	LOW	M/V	HIGH	M/V	M/V	HIGH	LOW		
Sulfate	Annual Geomean SOM	M/V	LOW	M/V	HIGH	LOW	HIGH	HIGH	M/V		
Temperature	Annual Geomean SOM	M/V	M/V	M/V	LOW	LOW	HIGH	HIGH	M/V		
Turbidity	Annual Geomean SOM	HIGH	M/V	M/V	M/V	M/V	HIGH	M/V	M/V		
Iron	Annual Geomean SOM	HIGH	M/V	M/V	LOW	LOW	HIGH	M/V	M/V		

Table 10 – Physical watershed variables considered for analysis in the LDA and SVM

Physical Watershed Variables	
Hydrologic Variables	Longest Flow Path, Network Density, Sum of Streams, Drainage Area, Slope %
Climatic Variables	Mean Annual Temperature, Mean Annual Precipitation/Mean Quarterly Precipitation
Ecoregion Variables	% Eastern Corn Belt, % Interior Plateau
Natural Region Variables	% Central Till Plain, % Bluegrass, % Highland Rim, % Shawnee Hills
Bedrock geology	%Gray Shale, % Limestone, %Limestone/Dolomite, %Sandstone/Limestone/Shale, %Siltstone
Point Sources	Confined Feeding Operations (CAFO/mi^2), Combined Sewer Overflows (CSO/$mi^{2)}$, National Pollution Discharge Elimination System sum of flow (NPDES/mi^2)
Land Use (2001)	%Urban, %Forest, %Cultivated Crops, %Grassland/Pasture/Scrubland, %Wetlands, %Water
Land Use Change (1991 – 2001)	%Urban to Forest, %Urban to Agriculture, %Agriculture to Urban, %Agriculture to Forest, %Forest to Urban, % Forest to Agriculture
Soil Drainage	Well to Excessively Drained, Moderately Well Drained, Somewhat Poorly Drained, Poor to Very Poorly Drained, Impervious Surface

Table 11 – Variable selection from stepwise LDA

Stepwise LDA Selected Variables	
Annual Mean	Interior Plateau, Drainage Area, Cultivated Crops, Shawnee Hills, Moderately Well Drained, CAFOs, Temperature, Forest to Urban, Water, Agriculture to Urban, Highland Rim
Annual Median	Highland Rim, NPDES, Cultivated Crops, Slope, Forest to Urban, Water, Precipitation, Agriculture to Urban, Grassland/Pasture/Scrubland, Wetlands, CAFOs, Agriculture to Forest, Sandstone/Limestone/Shale
Annual Trimmed Mean	Agriculture to Urban, Forest to Urban, Cultivated Crops, Moderately Well Drained, Sum of Streams
Annual Geometric Mean	Highland Rim, NPDES, Precipitation, Urban, Agriculture to Urban, Wetlands, Urban to Agriculture, Sandstone/Limestone/Shale, Eastern Corn Belt
Q1 Mean	Forest, Interior Plateau, Longest Flow Path, Temperature, Sandstone/Limestone/Shale, NPDES, Poorly Drained, Drainage Area, Urban to Agriculture, Gray Shale, Water, Network Density, Grassland/Pasture/Scrubland
Q1 Median	Forest, Interior Plateau, Longest Flow Path, Water, Forest to Urban, Network Density, Cultivated Crops, Agriculture to Forest, Shawnee Hills, Poorly Drained, CAFOs, CSOs, Temperature, Grassland/Pasture/Scrubland, Bluegrass, Gray Shale, Highland Rim
Q1 Trimmed Mean	Temperature, Interior Plateau, Cultivated Crops, Shawnee Hills, Forest, Forest to Urban, Agriculture to Forest, Central Till Plain, Gray Shale, Moderately Well Drained, NPDES, Drainage Area
Q1 Geometric Mean	Interior Plateau, Drainage Area, Wetlands, Forest, Shawnee Hills, Grassland/Pasture/Scrubland, Eastern Corn Belt, NPDES
Q2 Mean	Cultivated Crops, Interior Plateau, Sandstone/Limestone/Shale, Wetlands, Highland Rim, Agriculture to Urban, NPDES, Forest to Agriculture, Grassland/Pasture/Scrubland, Limestone, Forest to Urban
Q2 Median	Interior Plateau, Longest Flow Path, Cultivated Crops, Moderately Well Drained, Sandstone/Limestone/Shale, Highland Rim, Grassland/Pasture/Scrubland, Network Density, Forest to Urban, Water, Agriculture to Urban
Q2 Trimmed Mean	Forest, Drainage Area, Interior Plateau, Shawnee Hills, Wetlands, Longest Flow Path, NPDES, Urban, Highland Rim, CAFOs, Central Till Plain, CSOs, Agriculture to Forest, Precipitation, Urban to Forest, Gray Shale, Poorly Drained
Q2 Geometric Mean	Forest, NPDES, Highland Rim, Urban, Central Till Plain, Forest to Urban, Water, Limestone, Network Density, Sum of Streams, Grassland/Pasture/Scrubland, Agriculture to Forest, Temperature, Shawnee Hills, CSOs

Table 11 (cont.) – Variable selection from stepwise LDA

Stepwise LDA Selected Variables	
Q3 Mean	Drainage Area, Temperature, Urban, Agriculture to Forest, Forest to Urban, Water, Moderately Well Drained, Central Till Plain, Limestone
Q3 Median	Temperature, Longest Flow Path, CSOs, NPDES, Agriculture to Forest, Sum of Streams, Cultivated Crops, Poorly Drained, Wetlands
Q3 Trimmed Mean	Temperature, Urban, Bluegrass, Central Till Plain, Agriculture to Forest, Forest to Urban, Water, Network Density, Agriculture to Urban, Grassland/Pasture/Scrubland, Longest Flow Path, Poorly Drained
Q3 Geometric Mean	Interior Plateau, Longest Flow Path, Shawnee Hills, Cultivated Crops, Highland Rim, Moderately Well, Grassland/Pasture/Scrubland, Temperature, Forest to Urban, Bluegrass, Network Density, CAFOs, Eastern Corn Belt, Forest to Agriculture, Agriculture to Urban, Water, Limestone
Q4 Mean	Shawnee Hills, Interior Plateau, Central Till Plain, Urban, Forest to Urban, Forest, Moderately Well Drained, Highland Rim, NPDES
Q4 Median	NPDES, Highland Rim, Interior Plateau, Shawnee Hills, Cultivated Crops, Wetlands, Agriculture to Urban, Impervious Surface, Drainage Area
Q4 Trimmed Mean	Interior Plateau, Drainage Area, Shawnee Hills, Cultivated Crops, NPDES, Urban to Forest, Forest to Urban, Water
Q4 Geometric Mean	Forest, Drainage Area, Interior Plateau, Shawnee Hills, Highland Rim, NPDES, Moderately Well Drained, Urban

Table 12 – Pooled Covariance Matrix Rank; a test for Multicollinearity/Singularity

Dataset	# of Variables	Pooled Covariance Matrix Rank	Likely Multicollinearity/ Singularity Problems
Annual Mean	11	11	NO
Annual Median	13	13	NO
Annual Trimmed Mean	5	5	NO
Annual Geometric Mean	9	9	NO
Q1 Mean	13	13	NO
Q1 Median	17	17	NO
Q1 Trimmed Mean	12	12	NO
Q1 Geometric Mean	8	8	NO
Q2 Mean	11	11	NO
Q2 Median	11	11	NO
Q2 Trimmed Mean	17	17	NO
Q2 Geometric Mean	15	15	NO
Q3 Mean	9	9	NO
Q3 Median	9	9	NO
Q3 Trimmed Mean	12	12	NO
Q3 Geometric Mean	17	17	NO
Q4 Mean	9	9	NO
Q4 Median	9	9	NO
Q4 Trimmed Mean	8	8	NO
Q4 Geometric Mean	8	8	NO

Table 13 – Classification coefficients and constant for the annual geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-350.6	-245.8	-289.4	-281.3	-225.1
Highland Rim	139.4	102.7	132.0	102.9	101.2
NPDES	29.0	4.2	-22.5	-27.5	-2.0
Precipitation	236.3	215.4	263.8	269.5	200.2
Urban	170.7	155.8	159.7	170.0	121.7
Agriculture to Urban	-160.9	-134.7	-129.0	-118.8	-101.4
Wetlands	150.2	141.7	160.2	163.1	121.2
Urban to Agriculture	-41.0	-35.7	-47.3	-48.9	-25.2
Sandstone, Limestone, Shale	284.3	224.8	230.6	219.3	223.3
Eastern Corn Belt	391.6	335.6	344.4	345.0	327.9

Table 14 – Cross Validation Classification rates for the SVM and LDA

Datasets	SVM CV % Accuracy	LDA CV % Accuracy
Annual Mean	93.2	79.5
Annual Median	77.3	86.4
Annual Trimmed Mean	79.5	63.6
Annual Geometric Mean	88.6	77.3
Q1 Mean	72.7	77.3
Q1 Median	79.5	63.6
Q1 Trimmed Mean	81.8	70.5
Q1 Geometric Mean	84.1	86.4
Q2 Mean	65.9	86.4
Q2 Median	90.9	84.1
Q2 Trimmed Mean	84.1	90.9
Q2 Geometric Mean	70.5	90.9
Q3 Mean	79.5	68.2
Q3 Median	77.3	72.7
Q3 Trimmed Mean	86.4	81.8
Q3 Geometric Mean	70.5	88.6
Q4 Mean	75.0	79.5
Q4 Median	86.4	68.2
Q4 Trimmed Mean	75.0	61.4
Q4 Geometric Mean	75.0	70.5
AVERAGE	**79.7**	**77.4**

Table 15 – Hyperparameters and total number of support vectors for each SVM model

Data set	# of Support Vectors	C	γ
Annual Geometric Mean	42	16	0.125
Annual Mean	33	2	0.5
Annual Median	43	16	0.125
Annual Trimmed Mean	40	16	0.125
Q1 Geometric Mean	43	8	0.125
Q1 Mean	40	128	0.0625
Q1 Median	38	32	0.0625
Q1 Trimmed Mean	37	8	0.25
Q2 Geometric Mean	41	16	0.125
Q2 Mean	42	16	0.0625
Q2 Median	44	8	0.125
Q2 Trimmed Mean	41	2	1
Q3 Geometric Mean	43	2	0.5
Q3 Mean	42	1	0.5
Q3 Median	44	16	0.125
Q3 Trimmed Mean	35	2	0.5
Q4 Geometric Mean	42	4	1
Q4 Mean	43	2	1
Q4 Median	40	8	0.5
Q4 Trimmed Mean	41	16	0.5

Table 16 – Misclassified stations after resubstitution. The first number in the parentheses indicates the cluster each station was misclassified from and the second number represents the cluster into which they were classified

Dataset	LDA Misclassified Stations	SVM Misclassified Stations
Annual Mean	WR-279 (6, 2), WR-319 (1, 2)	EW-168 (2, 1), WR-279 (3, 1)
Annual Median	WR-279 (3, 1)	NONE
Annual Trimmed Mean	EC-7 (4, 2), EW-79 (3, 1), FR-64 (4, 1), IN-2 (3, 5), MC-35 (4, 5), MU-20 (4, 3), SGR-1 (5, 4), SND-4 (5, 1), VF-38 (3, 5), WR-134 (5, 3), WR-293 (1, 5), WR-309 (1, 3), WR-46 (1,3)	NONE
Annual Geometric Mean	CIC-17 (2,5), EC-21 (2, 5), SND-4 (3, 4)	WR-192 (7, 4), WR-210 (7, 4)
Q1 Mean	FC-26 (2, 6)	CIC-17, EW-1, EW-168, EW-94
Q1 Median	NONE	WR-134
Q1 Trimmed Mean	FC-7 (7, 5), FR-64 (7, 1), WR-319 (7, 5), WR-348 (1, 7)	FC-7
Q1 Geometric Mean	BL-.7 (2, 5), EC-7 (4, 5), EW-168 (1, 2)	EC-1 (6, 7), EC-7 (6, 7), FC-0.6 (6, 7), FC-7 (4,7), WR-319 (4, 7)
Q2 Mean	EW-79 (5, 2), FC-26 (3, 1)	BL-.7 (8, 5), EW-168 (5, 8), EW-79 (4, 7), FR-17 (8, 5), FR-64 (8, 5), WR-319 (5, 3), WR-348 (8, 5)
Q2 Median	EW-79 (1, 4)	NONE
Q2 Trimmed Mean	EW-239 (6, 7)	NONE
Q2 Geometric Mean	NONE	BWC-4 (2, 3), EW-168 (2, 3), FC-7 (2, 6), MC-18 (2, 3), WR-192 (9, 8), WR-319 (2, 3)
Q3 Mean	BL-.7 (1, 4), CIC-17 (2, 4), EC-21 (2, 1), WR-162 (6, 2)	EC-7 (2, 3)
Q3 Median	EC-21 (1, 4), FC-0.6 (3, 1), FC-26 (2, 1), FC-7 (3, 1)	EC-1 (4, 2)
Q3 Trimmed Mean	EC-21 (5, 3), WR-319 (3, 5)	EC-7 (3, 1), FC-7(3, 1), WR-293 (4, 1), WR-309 (1, 4)
Q3 Geometric Mean	NONE	CIC-17 (3, 7), EC-21 (3, 7), EEL-1 (4, 6), EW-1 (4, 6), EW-94 (4, 6), WR-248 (2, 3), WR-309 (3, 7)
Q4 Mean	EC-21 (5, 1), EW-79 (3, 2), FC-26 (1, 5), WR-279 (1, 5), WR-293 (1, 5), WR-309 (5, 1)	WR-134 (1, 5), WR-309 (1, 3)
Q4 Median	BL-.7 (2, 7), CIC-17 (7, 2), EC-1 (4, 7), EC-21 (7, 2), EW-239 (2, 5), WR-248 (4, 2)	EC-7 (4, 5)
Q4 Trimmed Mean	BL-64 (2, 5), EW-168 (3, 1), FC-0.6 (5, 2), FC-26 (1, 2), FR-64 (3, 1), SND-4 (3, 1), WR-279 (1, 2), WR-309 (2, 3), WR-319 (3, 1)	NONE
Q4 Geometric Mean	BL-64 (3, 1), CIC-17 (1, 5), EC-21 (3, 5), SGR-1 (5, 1), WR-319 (1, 3)	WR-81 (3, 2)

Table 17 – Cluster prediction and posterior probability error rate estimates for the Annual LDA models' classification of the ECWMP sites

ANNUAL	Geometric Mean (6)		Mean (5)		Median (5)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	2	0.386	5	0.985	3	1	4	0.618
ECWMP-02	5	1	3	0.999	5	0.62	2	0.887
ECWMP-03	5	0.957	2	0.793	1	1	4	0.562
ECWMP-04	5	0.995	3	1	2	1	2	0.991
ECWMP-05	5	1	1	0.999	1	1	4	0.65
ECWMP-06	5	0.969	5	0.999	3	0.67	4	0.327
ECWMP-07	5	0.973	5	1	3	1	1	0.431
ECWMP-08	2	0.695	1	0.975	1	1	1	0.379
ECWMP-09	5	0.998	5	1	3	1	1	0.493
ECWMP-10	5	0.998	5	1	3	1	1	0.386
ECWMP-11	5	0.995	5	1	3	1	1	0.438

Table 18 – Cluster prediction and probability estimates for the Annual SVM models' classification of the ECWMP sites

ANNUAL	Geometric Mean(8)		Mean (3)		Median (7)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	5	0.431	1	0.882	2	0.436	4	0.451
ECWMP-02	5	0.383	1	0.918	4	0.392	6	0.402
ECWMP-03	4	0.496	1	0.934	4	0.535	6	0.556
ECWMP-04	4	0.473	1	0.929	4	0.506	6	0.529
ECWMP-05	5	0.438	1	0.907	2	0.445	4	0.456
ECWMP-06	5	0.565	1	0.928	2	0.572	4	0.589
ECWMP-07	5	0.659	1	0.871	2	0.666	4	0.692
ECWMP-08	4	0.437	1	0.921	4	0.464	6	0.48
ECWMP-09	5	0.673	1	0.85	2	0.679	4	0.706
ECWMP-10	5	0.687	1	0.836	2	0.693	4	0.718
ECWMP-11	5	0.656	1	0.844	2	0.662	4	0.688

Table 19 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the annual datasets. Highlighted values classified the highest percentage of variables within the specified range among different models for a given station

	Annual SVM Model Cluster Range Accuracy				Annual LDA Model Cluster Range Accuracy			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	78.6	71.4	85.7	85.7	64.3	57.1	85.7	78.6
ECWMP-02	71.4	92.9	78.6	78.6	71.4	85.7	71.4	35.7
ECWMP-03	78.6	71.4	78.6	64.3	64.3	78.6	78.6	85.7
ECWMP-04	78.6	85.7	85.7	78.6	78.6	71.4	57.1	28.6
ECWMP-05	64.3	71.4	78.6	57.1	71.4	14.3	85.7	71.4
ECWMP-06	78.6	71.4	78.6	78.6	78.6	57.1	78.6	71.4
ECWMP-07	78.6	71.4	71.4	85.7	78.6	35.7	78.6	71.4
ECWMP-08	85.7	71.4	71.4	78.6	85.7	50.0	71.4	78.6
ECWMP-09	85.7	57.1	71.4	85.7	85.7	35.7	85.7	71.4
ECWMP-10	78.6	57.1	85.7	71.4	78.6	50.0	85.7	64.3
ECWMP-11	64.3	57.1	64.3	64.3	57.1	50.0	71.4	57.1

Table 20 – Wilcoxon Matched-Pairs Signed-Ranks test decision matrix; the SVM outperformed the LDA in 6/20 models; the LDA outperformed the SVM in 2/20 models; neither the LDA or SVM outperformed one another in 12/20 models

	Geometric Mean	Mean	Median	Trimmed Mean
Annual	EITHER	SVM	EITHER	EITHER
Quarter 1	EITHER	SVM	EITHER	SVM
Quarter 2	EITHER	LDA	LDA	SVM
Quarter 3	EITHER	SVM	EITHER	SVM
Quarter 4	EITHER	EITHER	EITHER	EITHER

FIGURES

Figure 1 – White River watershed 8 – digit HUCs; Upper White (05120201), Lower White (05120202), Eel (05120203), Driftwood (05120204), Flatrock –Haw (05120205), Upper East Fork (05120206), Muscatatuck (05120207), and Lower East Fork (05120208)

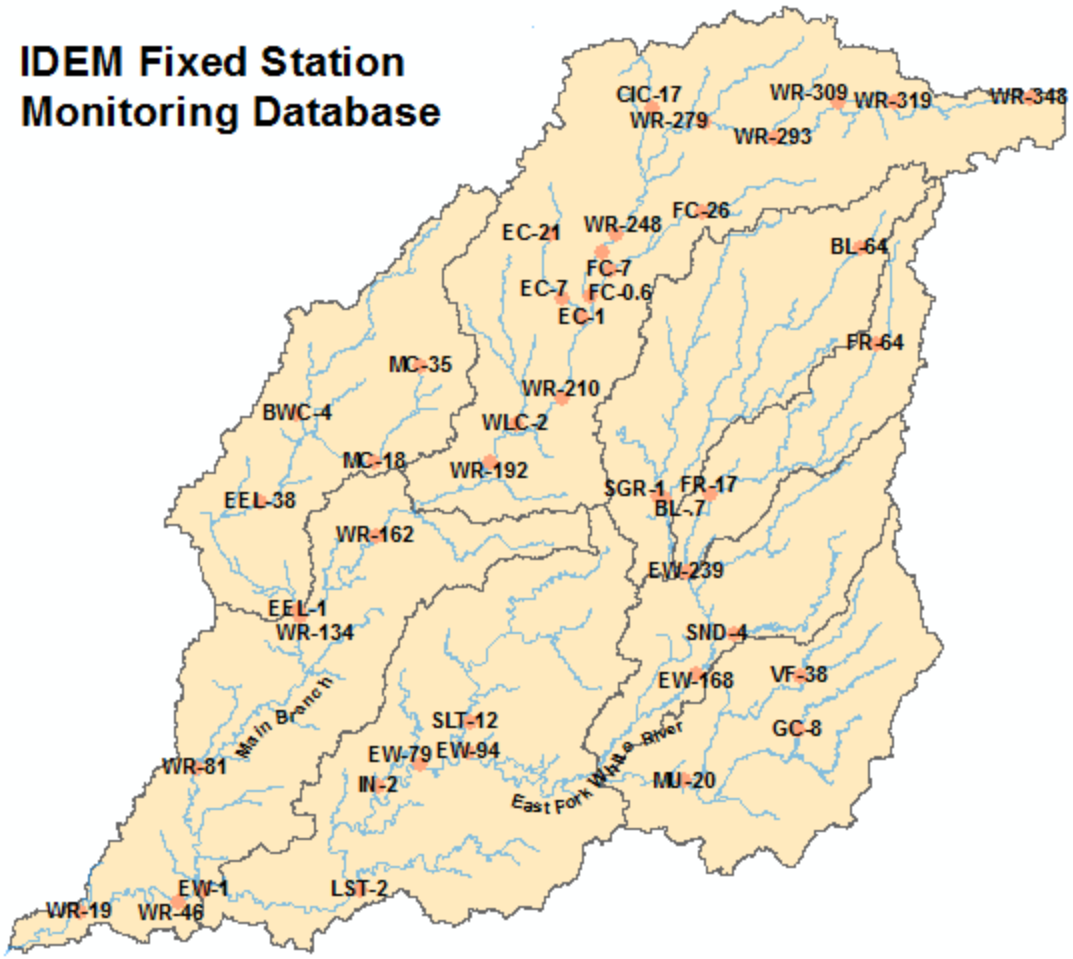


Figure 2 – Selected sites from the IDEM Fixed Station Monitoring network throughout the White River Watershed

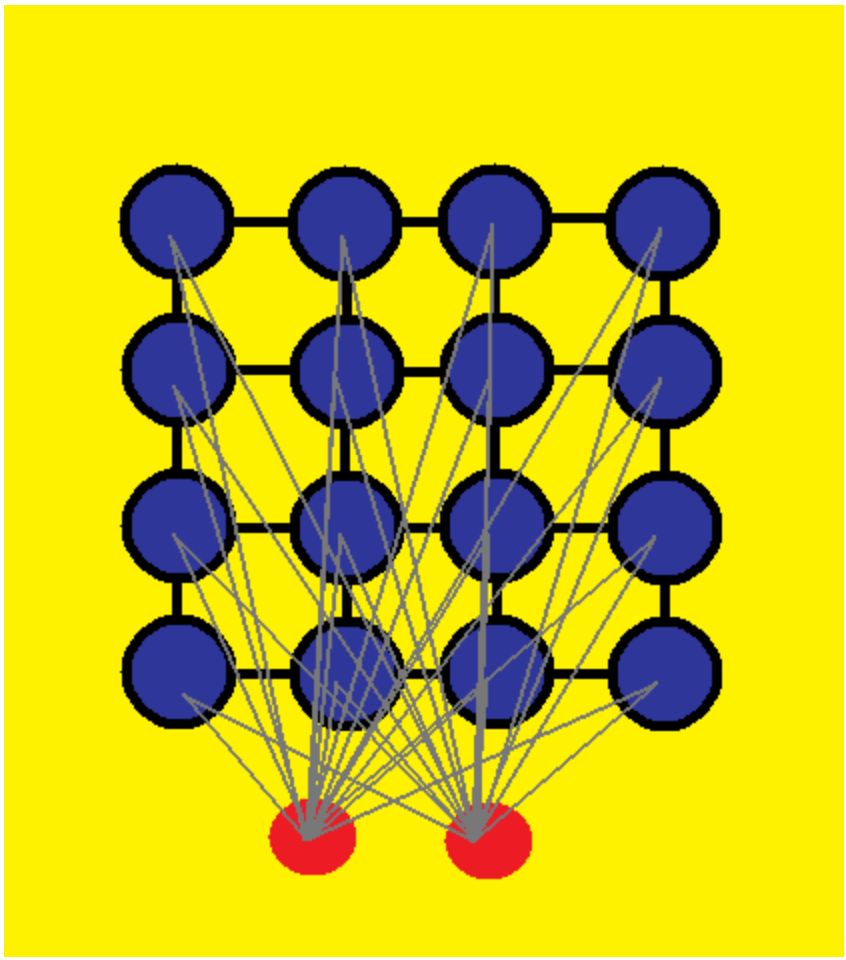


Figure 3 – Simplified Kohonen Self – Organizing Map; this example exhibits a 4 x 4 architecture and a rectangular topology; the smaller red circles represent the input nodes and the larger blue circles represent the output nodes

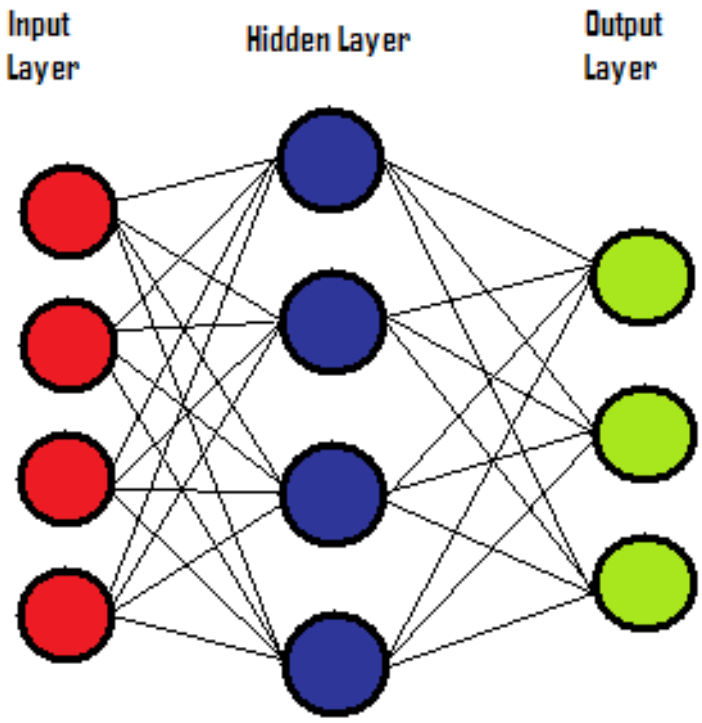


Figure 4 – A simplified representation of the architecture of a Support Vector Machine and an Artificial Neural Network

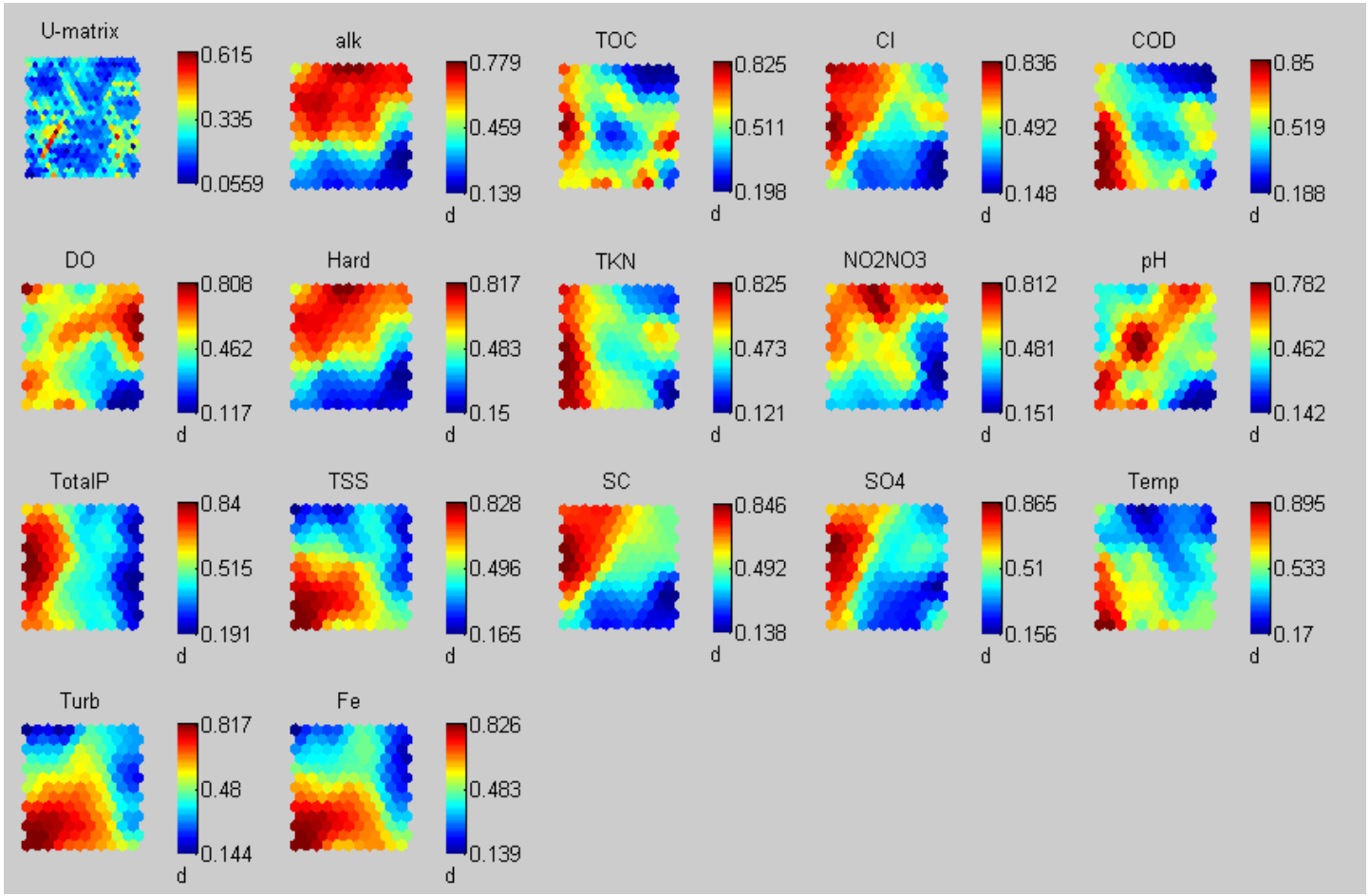


Figure 5 - Annual Mean Dataset Component Maps

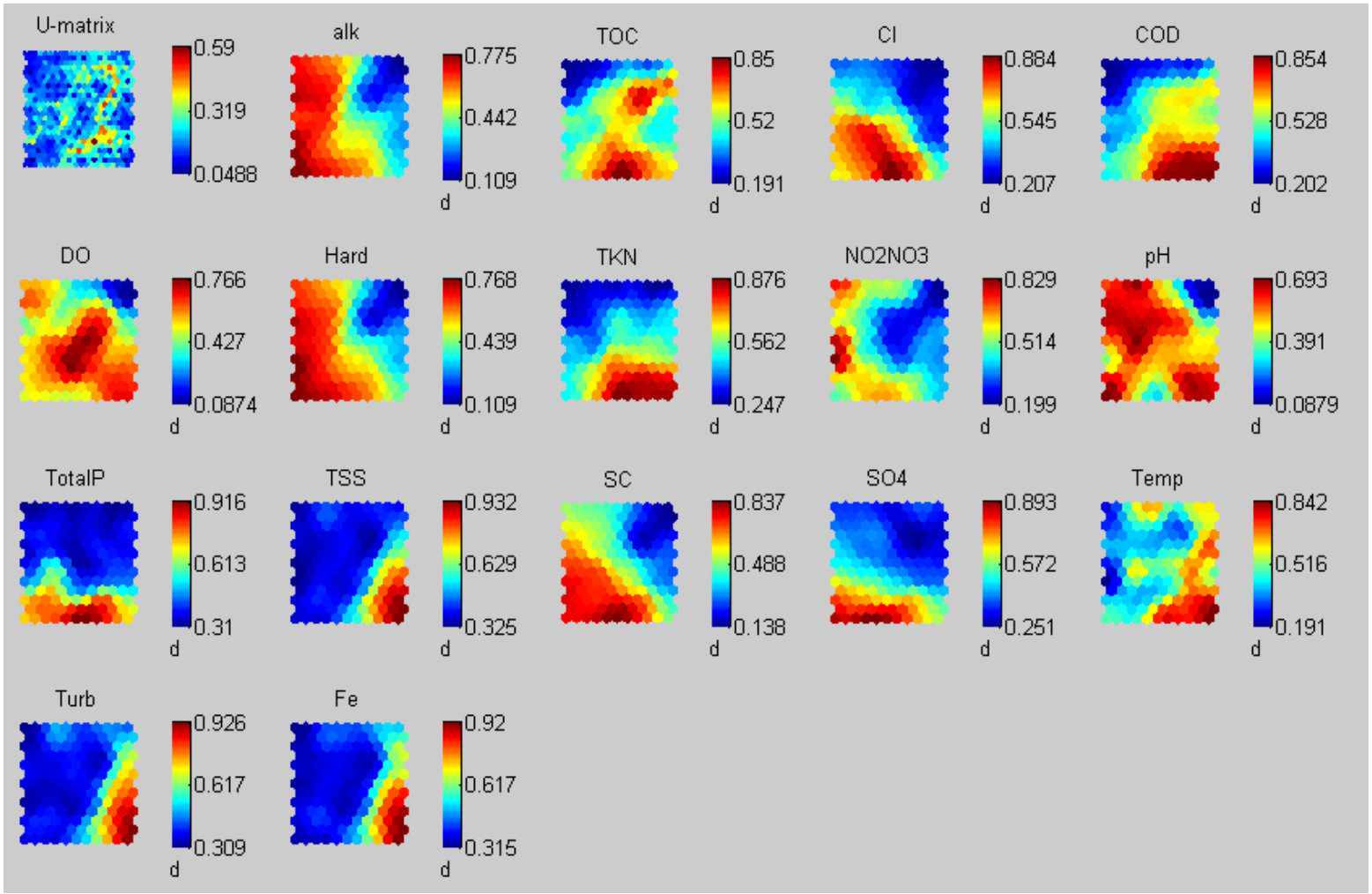


Figure 6 - Annual Median Dataset Component Maps

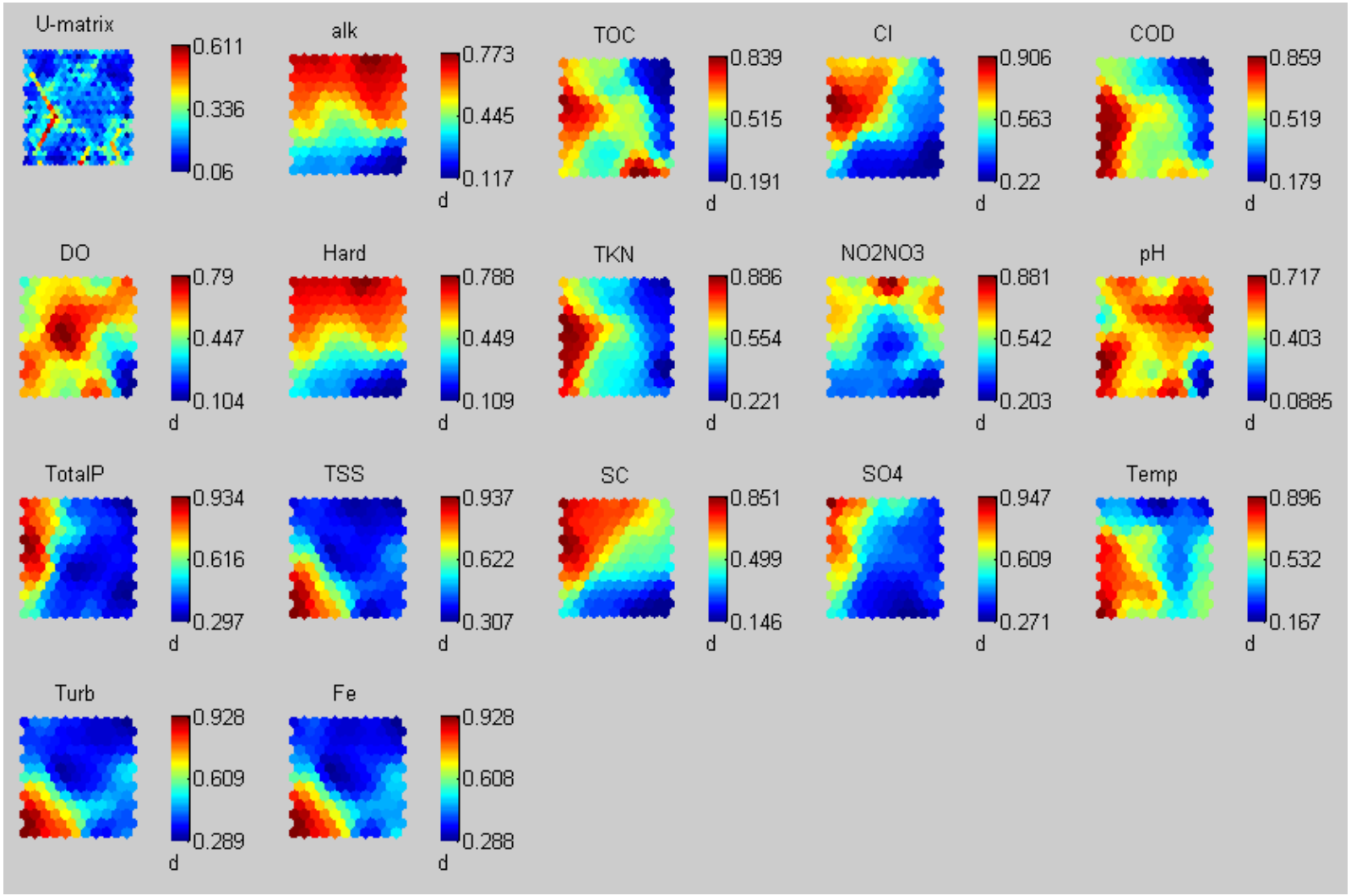


Figure 7 - Annual Trimmed Mean Dataset Component Maps

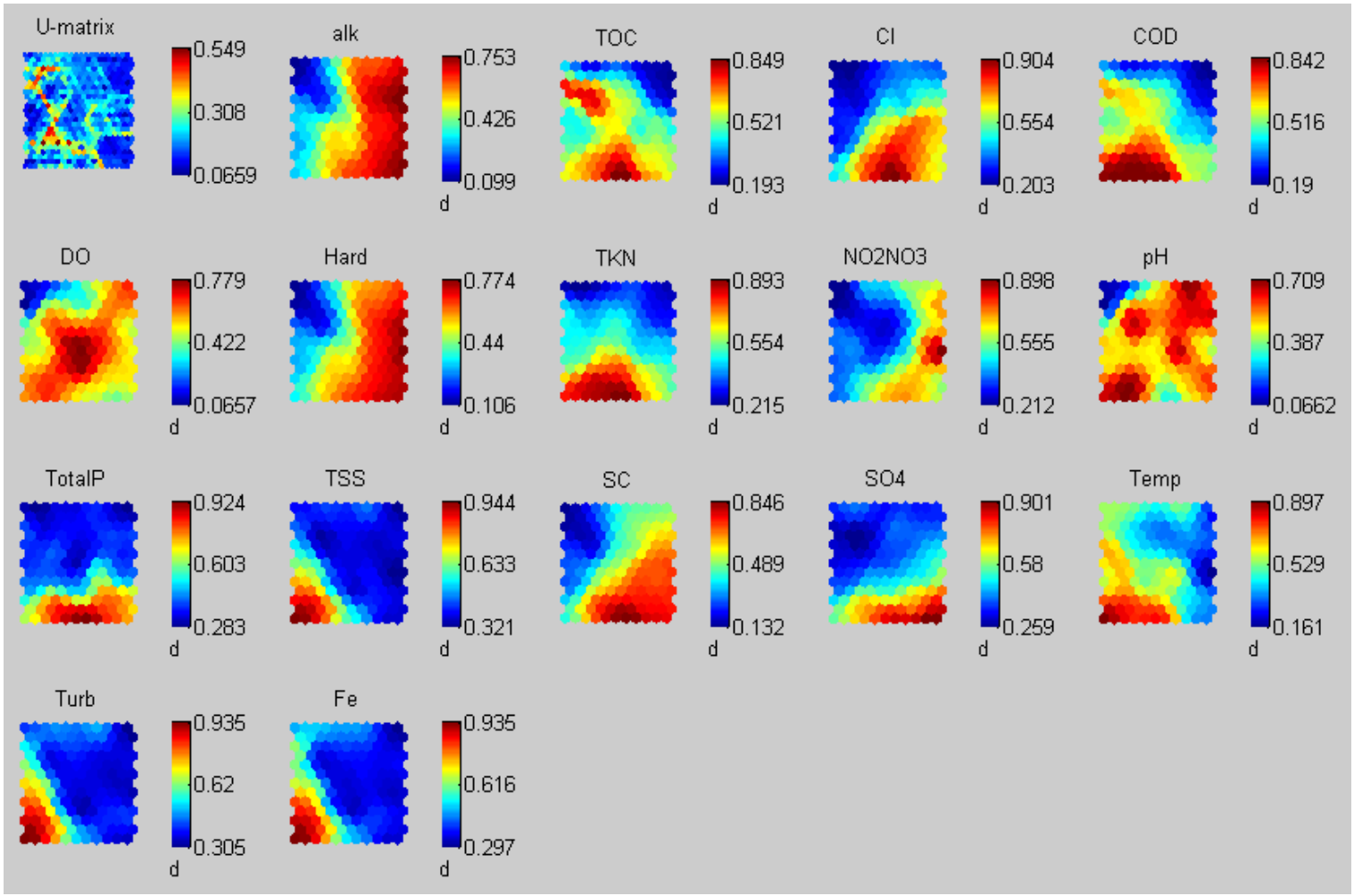


Figure 8 - Annual Geometric Mean Dataset Component Maps

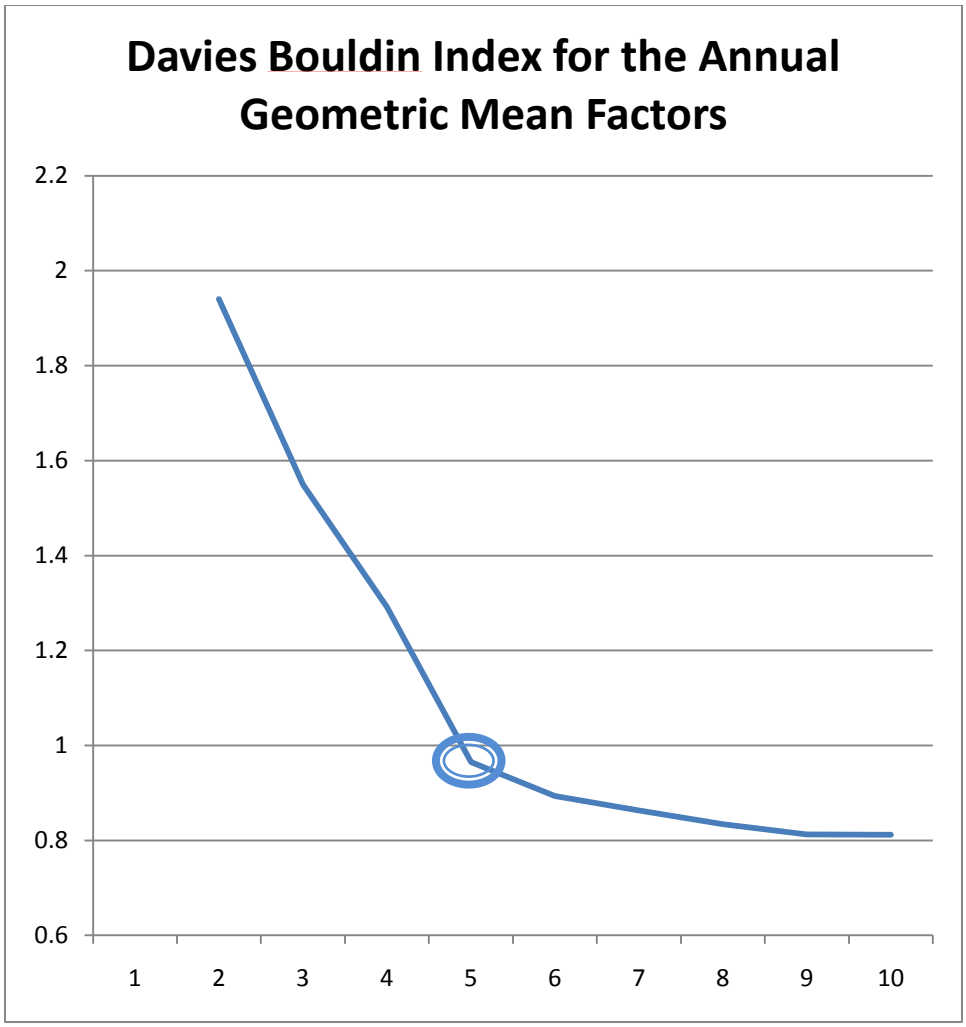


Figure 9 – Davies Bouldin index falls sharply until 5 clusters, then the slope levels out; therefore, 5 clusters was chosen to represent the annual geometric mean factor clusters

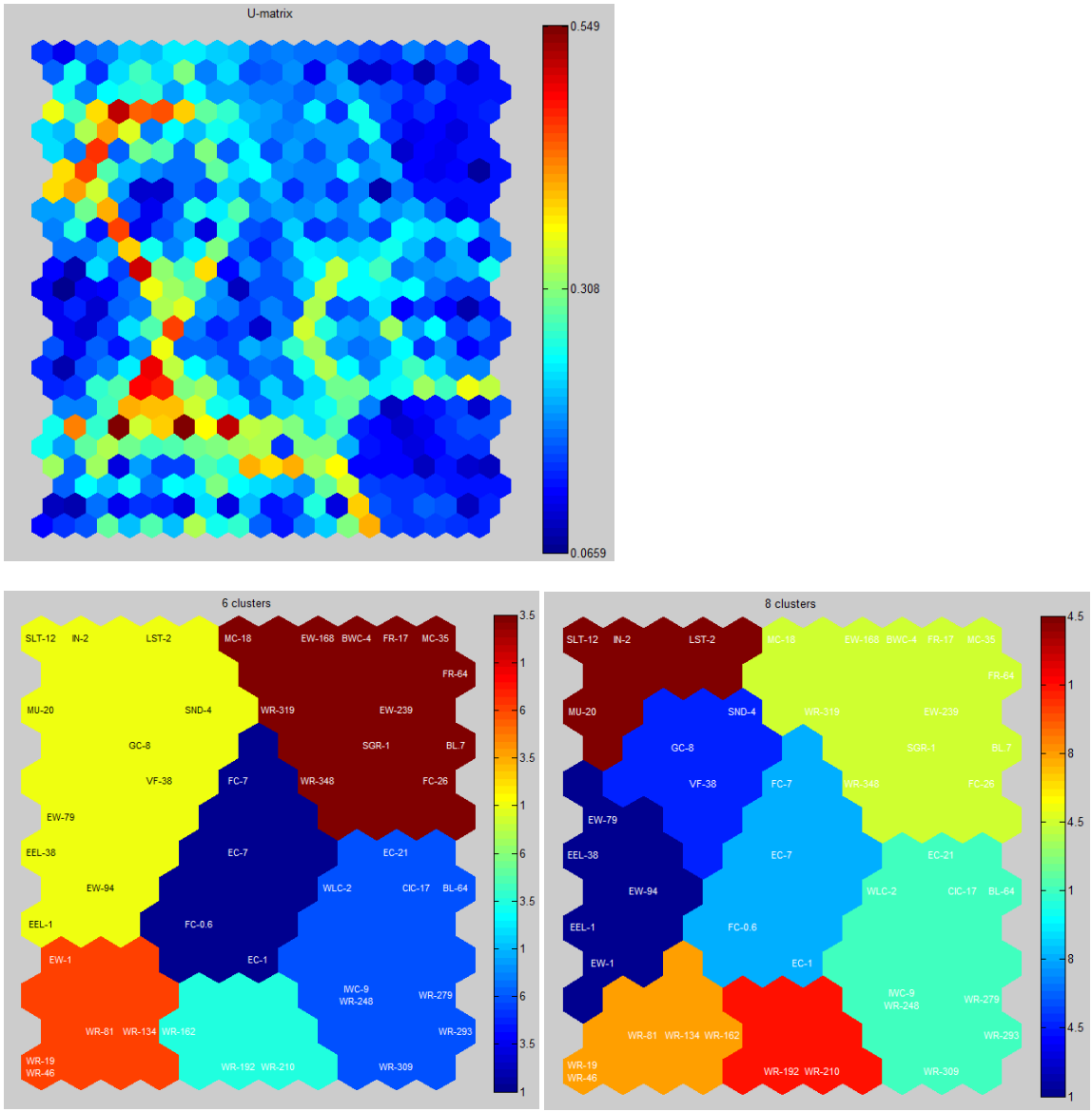


Figure 10 – The U-matrix and cluster arrangements of 6 and 8 clusters for the annual geometric mean; 8 clusters were chosen because they were better in line with the U-matrix

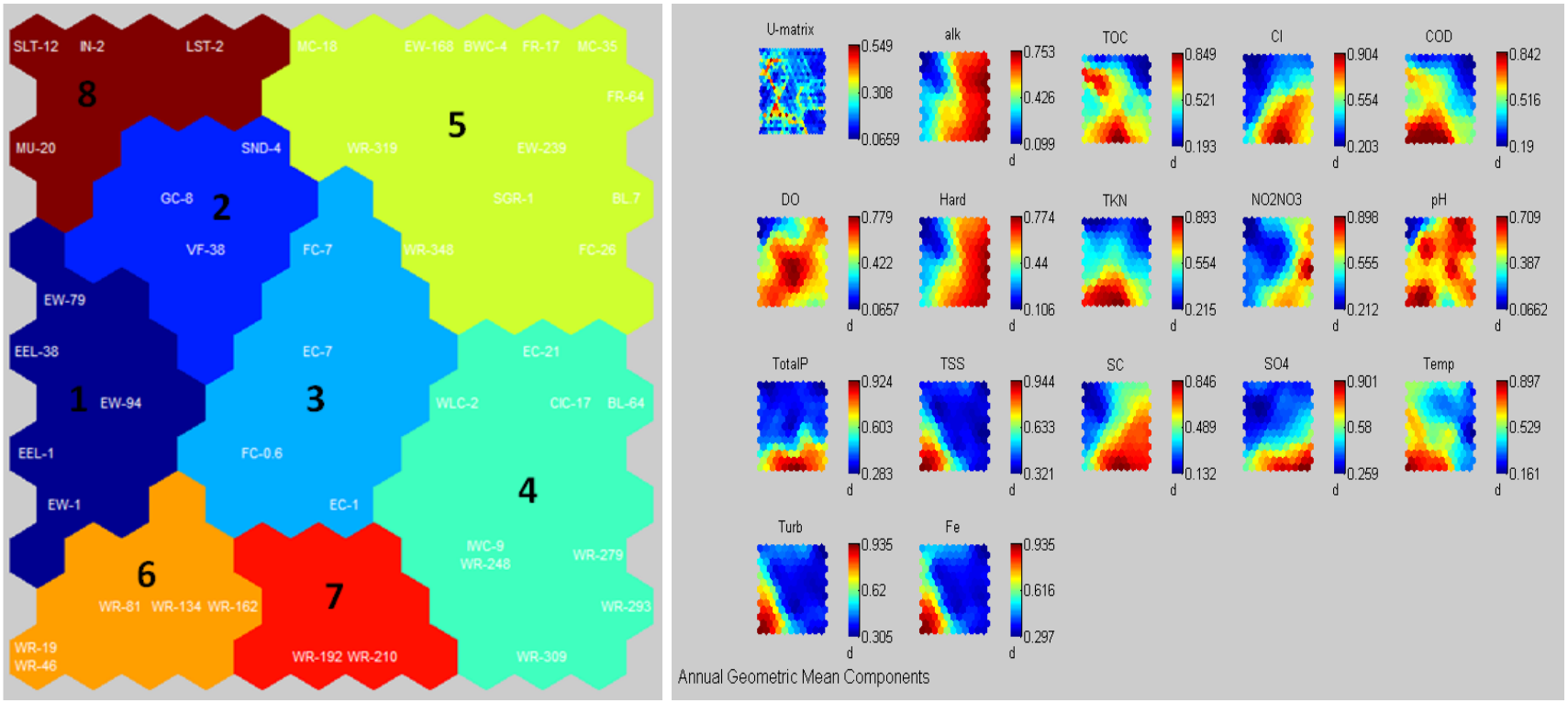


Figure 11 – Side by side comparison of the annual geometric mean SOM cluster configuration and the corresponding component maps; by visually overlaying the clustering arrangement figure on each of the component maps, one can identify variables with high values for each of the given clusters

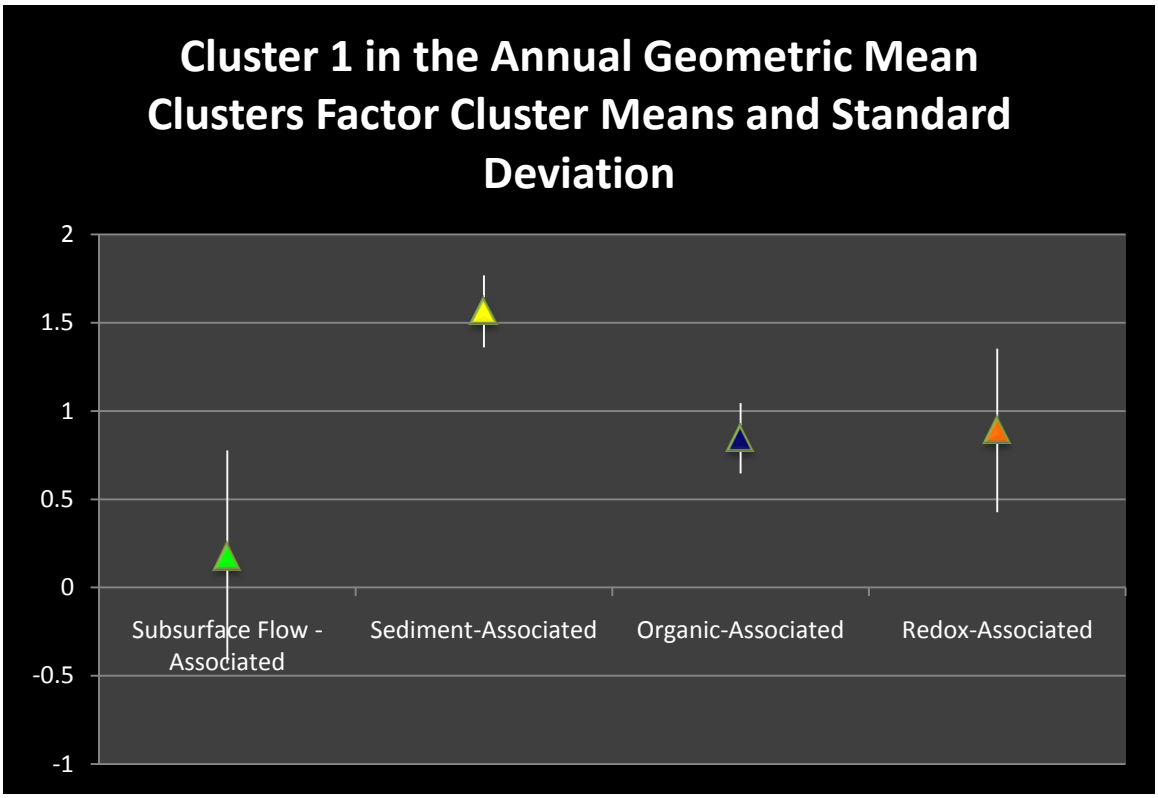


Figure 12 – The Box-Plot of cluster 1 from the annual geometric mean factor clusters; this cluster has moderate to variable concentrations of the subsurface flow related variables, high concentrations of the sediment relate variables, slightly high concentrations of the organic related variables, and high but variable values for the redox associated variables

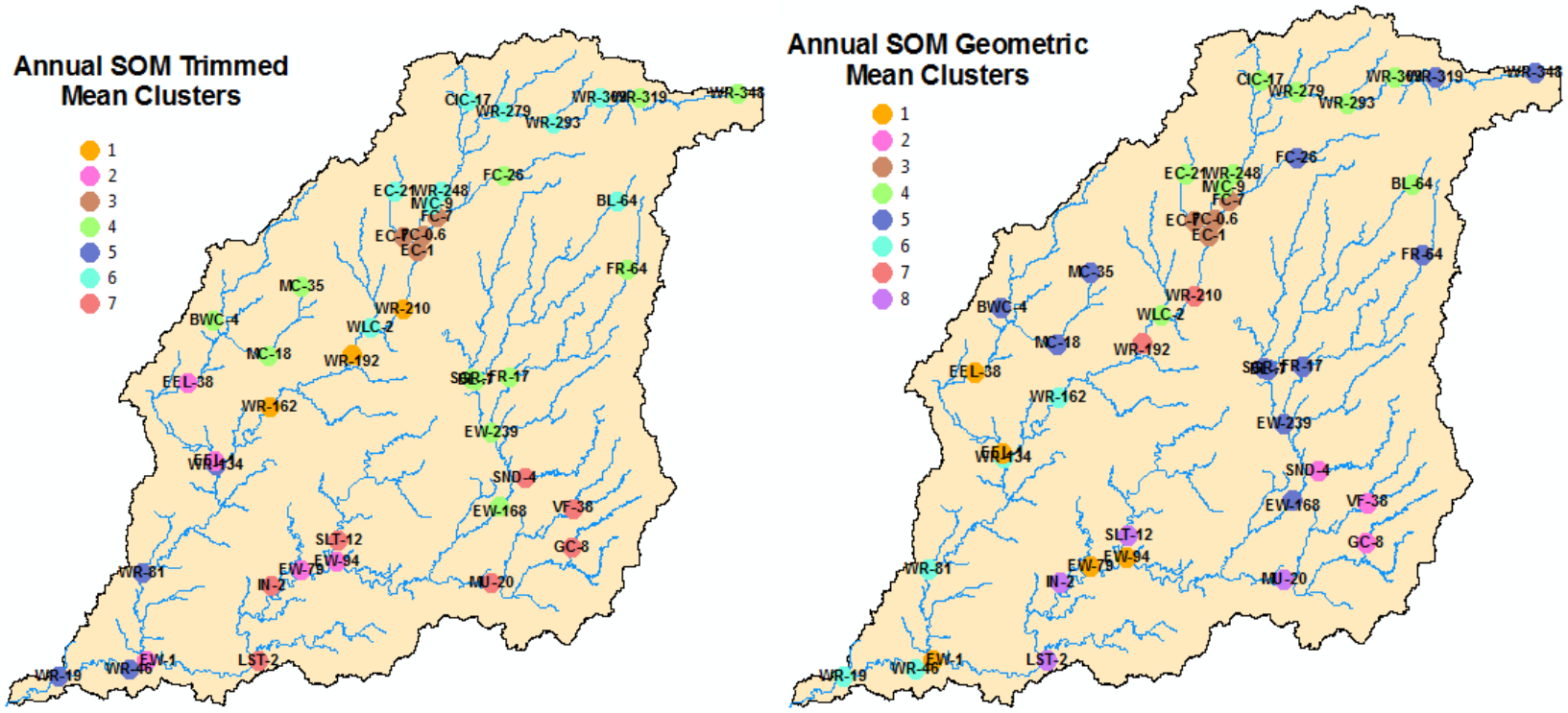


Figure 13 – Side by side comparison of the annual geometric mean SOM clusters and the annual trimmed mean SOM clusters; clusters 2 and 8 from the geometric mean clustering become cluster 7 in the trimmed mean clustering

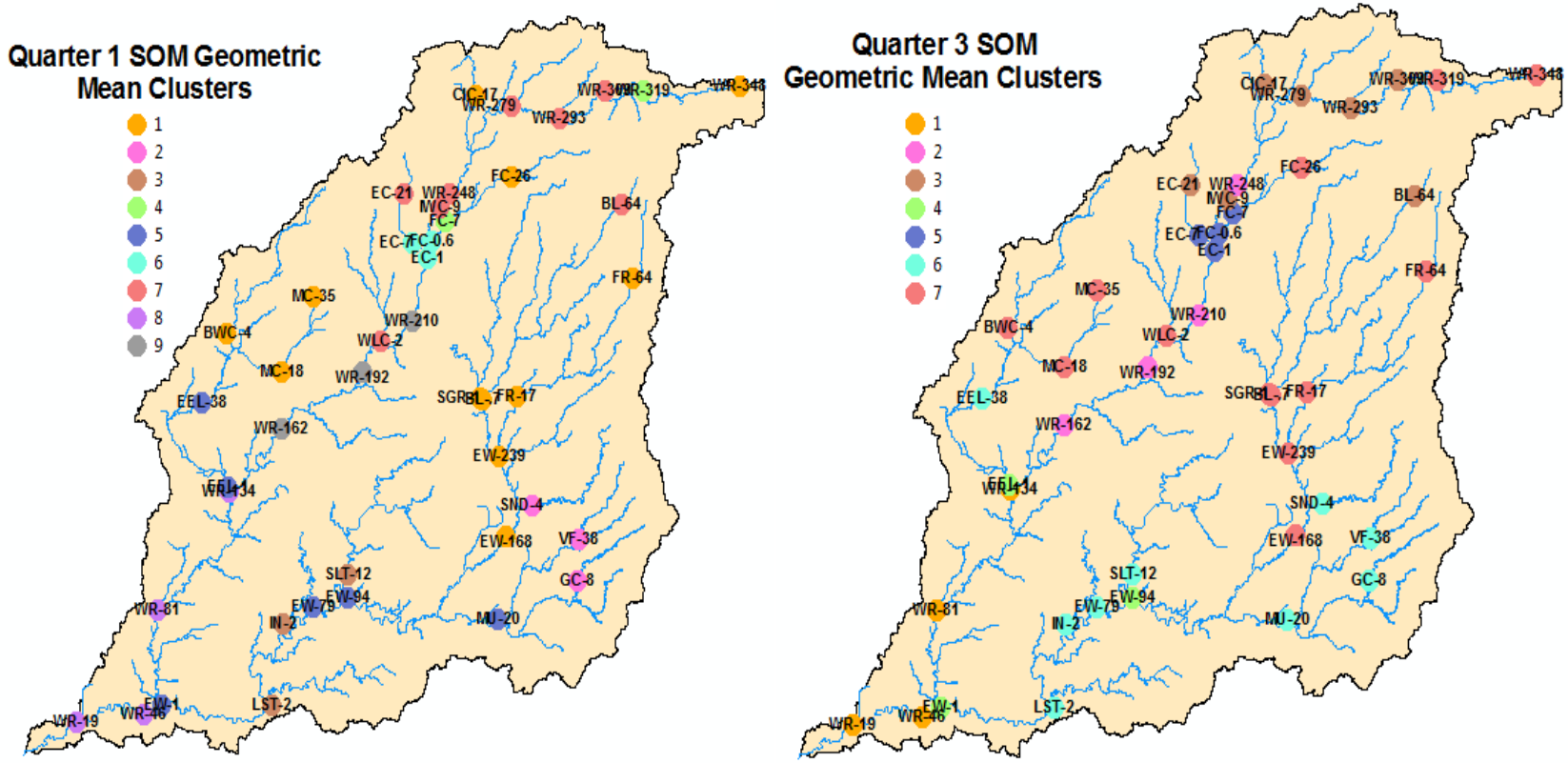


Figure 14 – Side by side comparison of the quarter 1 geometric mean SOM clusters and the quarter 3 geometric mean SOM clusters; CIC-17 and EW-239 are in the same cluster in quarter 1 and different clusters in quarter 3. In quarter 1 they are in cluster 1, which is characterized by low total phosphorus concentrations; however, in quarter 3 CIC-17 is in cluster 3, which is characterized by high total phosphorus concentrations, and EW-239 is in cluster 7, which is characterized by moderate total phosphorus concentrations

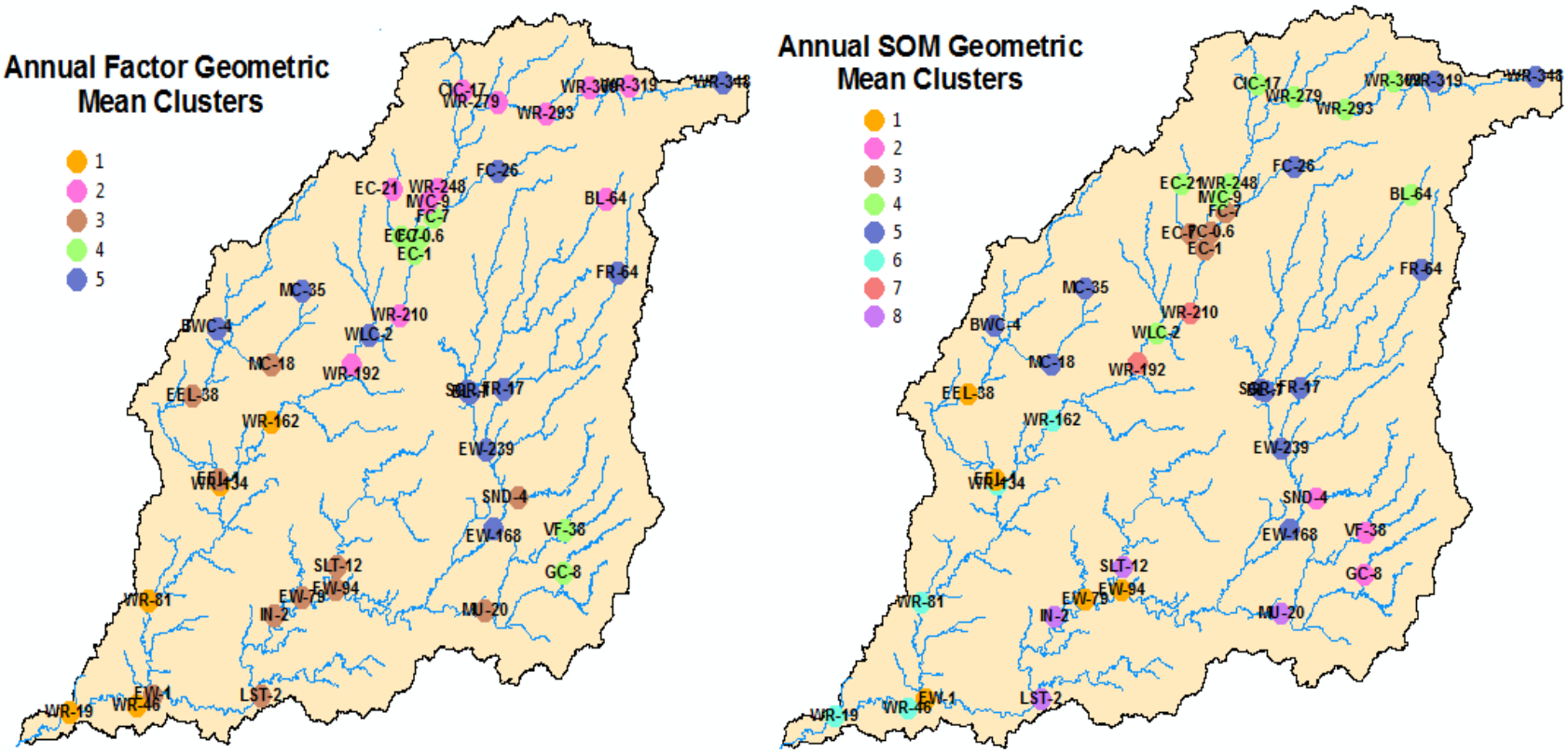


Figure 15 – Side by side comparison of the annual geometric mean factor clusters and the annual geometric mean SOM clusters; the clustering of the SOM and factors produce similar clusters, but there are some differences (e.g. SND-4)

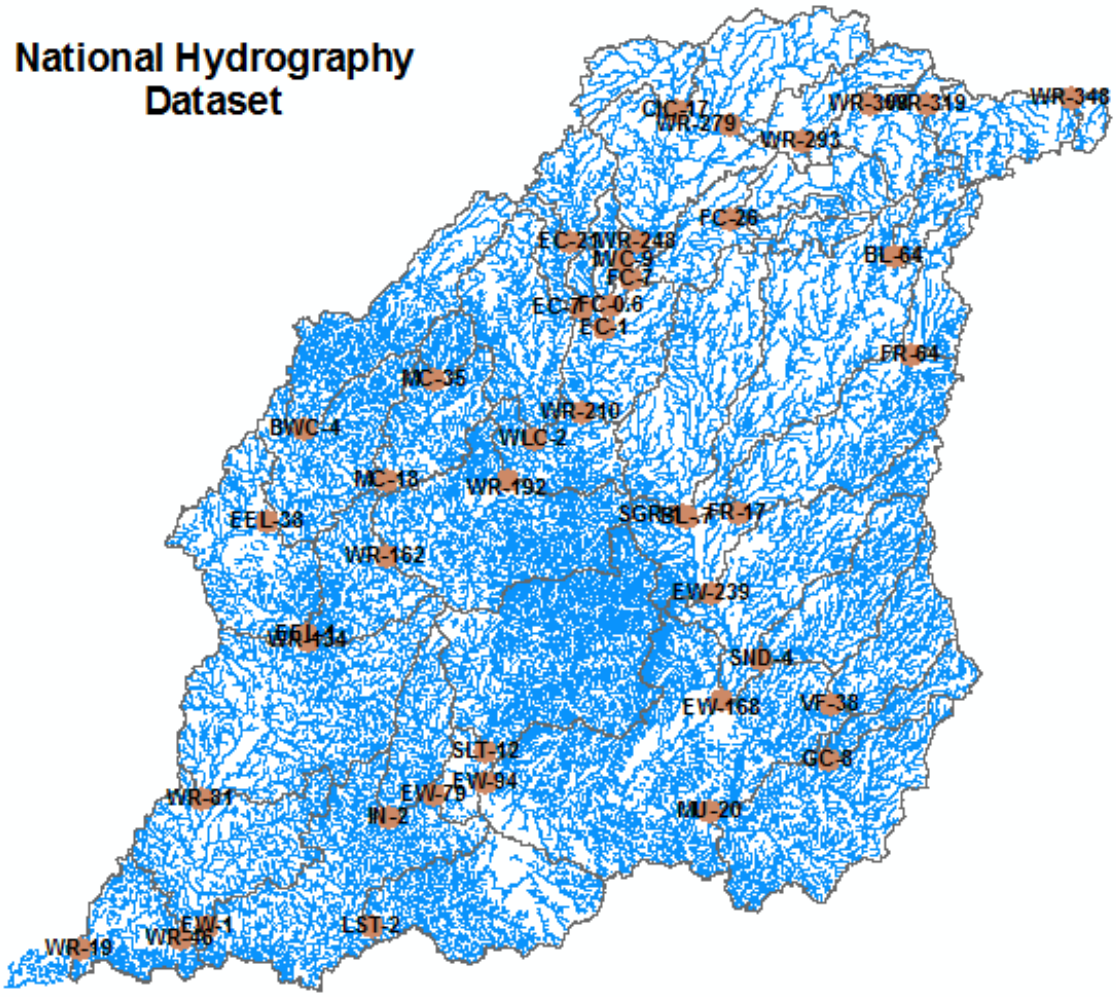


Figure 16 – Individual monitoring station watersheds and the National Hydrography Dataset; it is visually apparent that network density increases in the southern half of the watershed

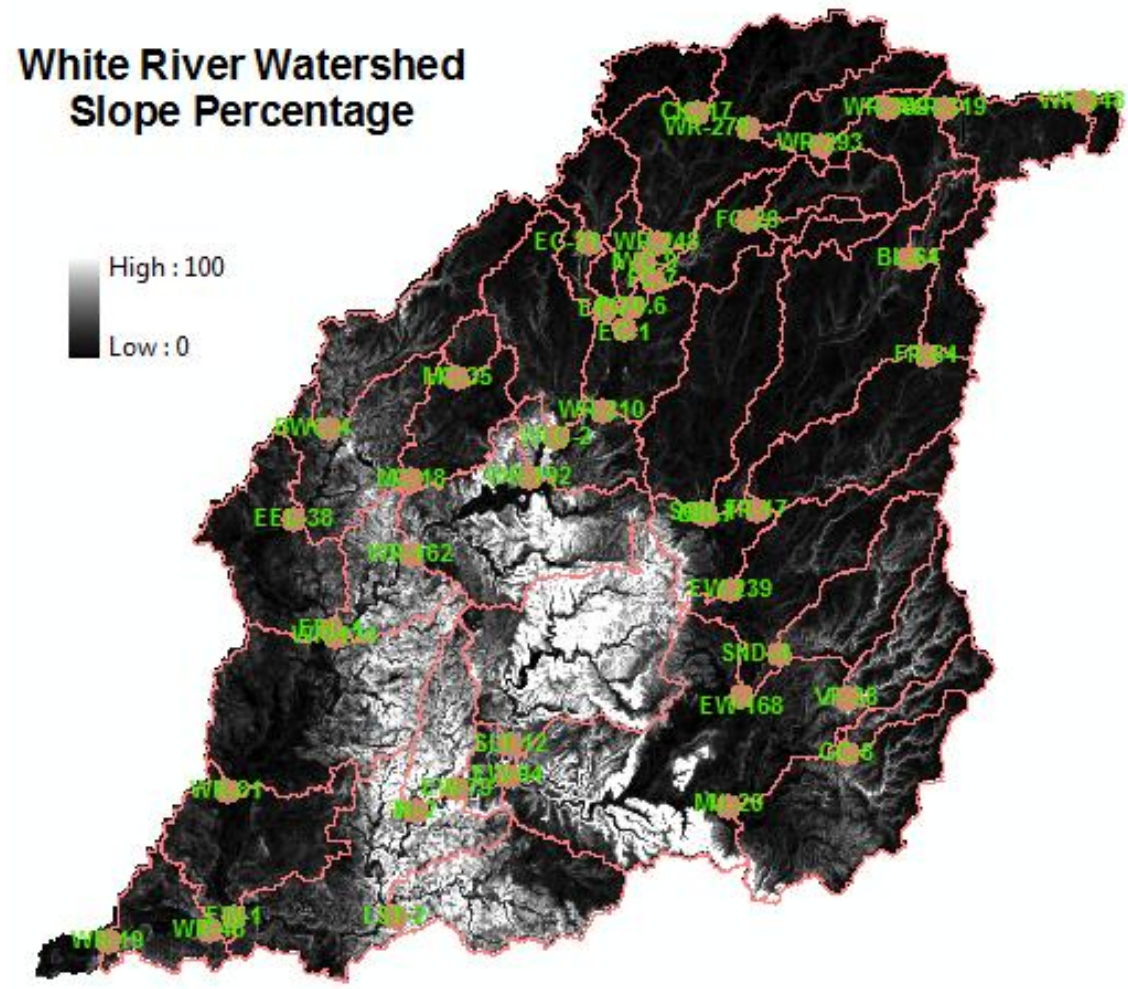


Figure 17 – The slope percentages of the White River watershed. Flat areas are darker and steeper slopes are lighter

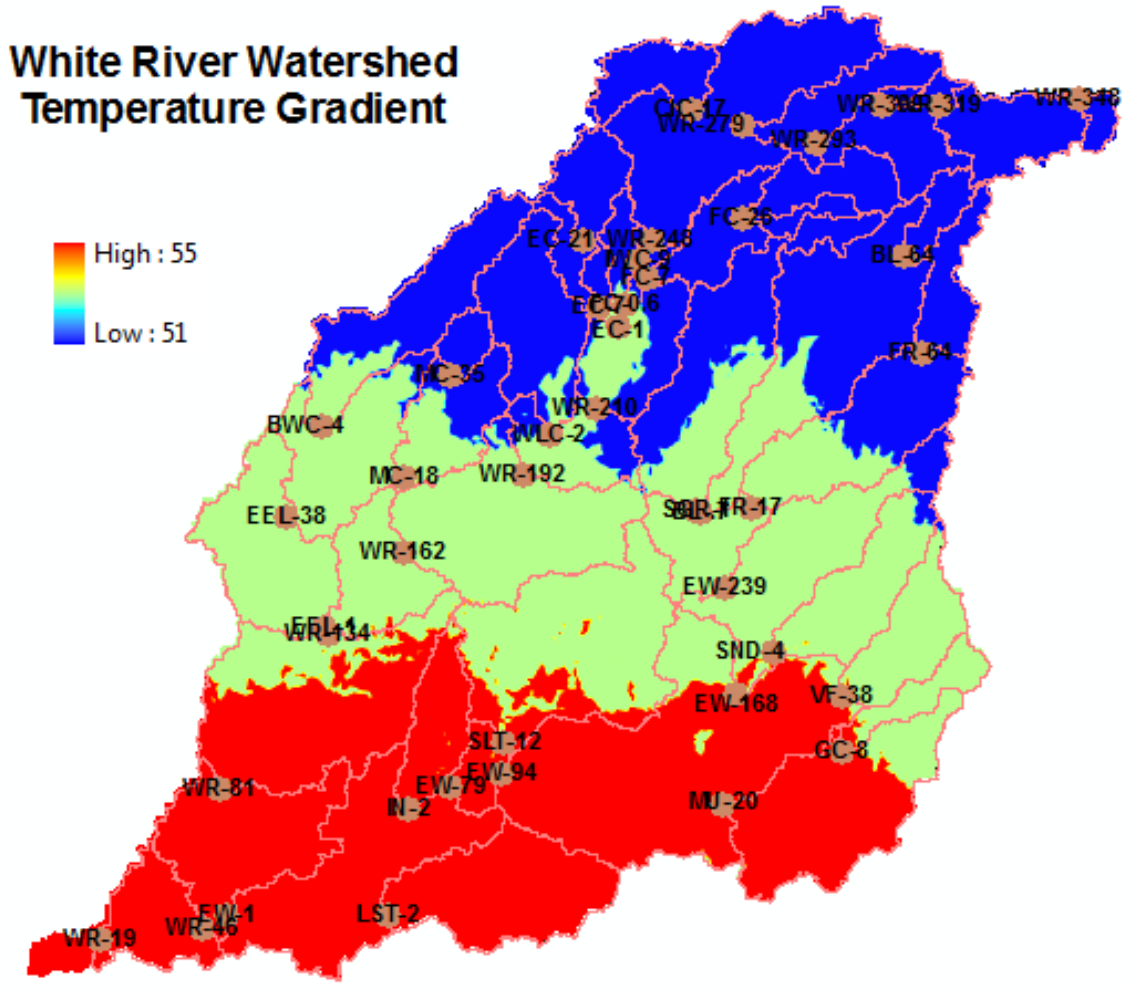


Figure 18 – White River Watershed mean annual temperature gradient (values are in degrees Celsius)

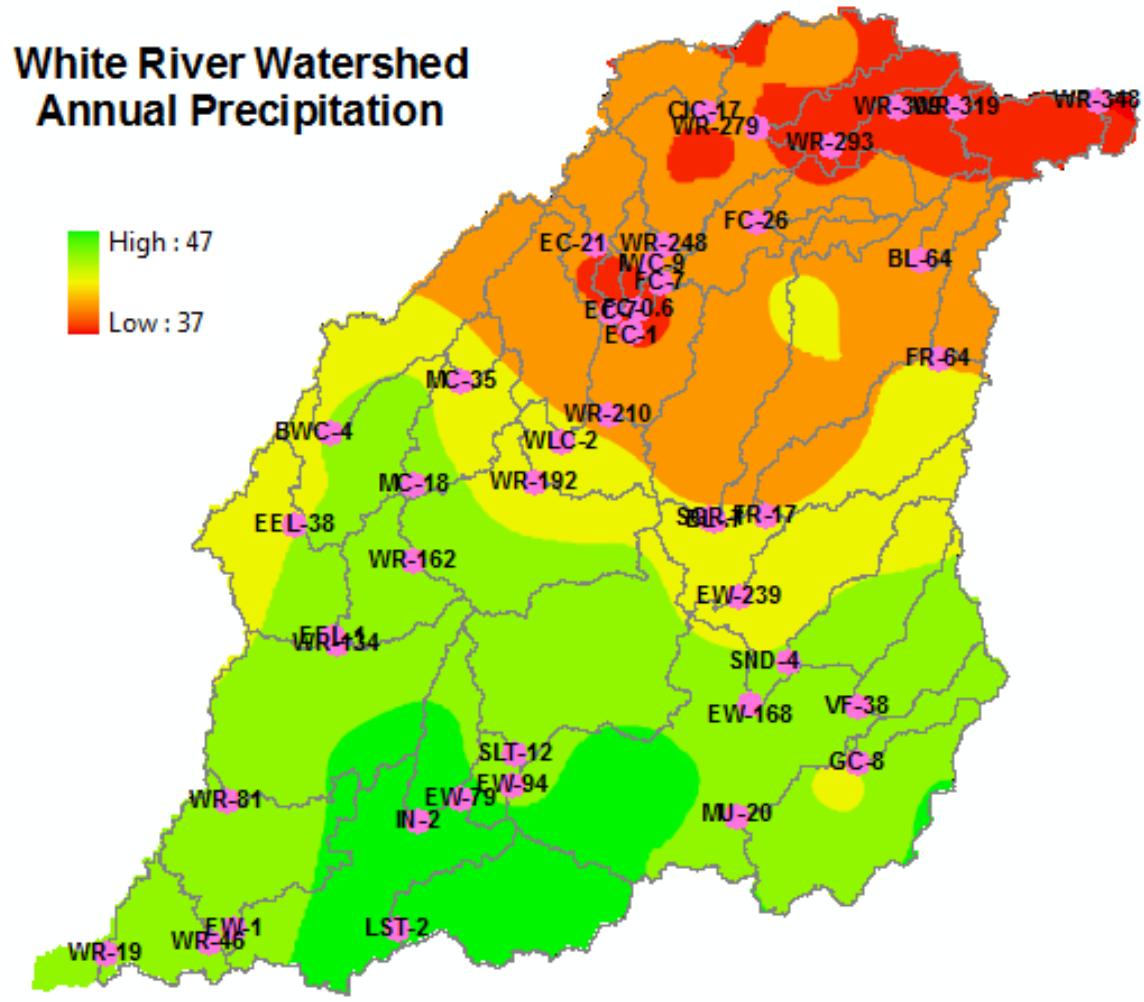


Figure 19 – White River Watershed mean annual precipitation (values are in inches)

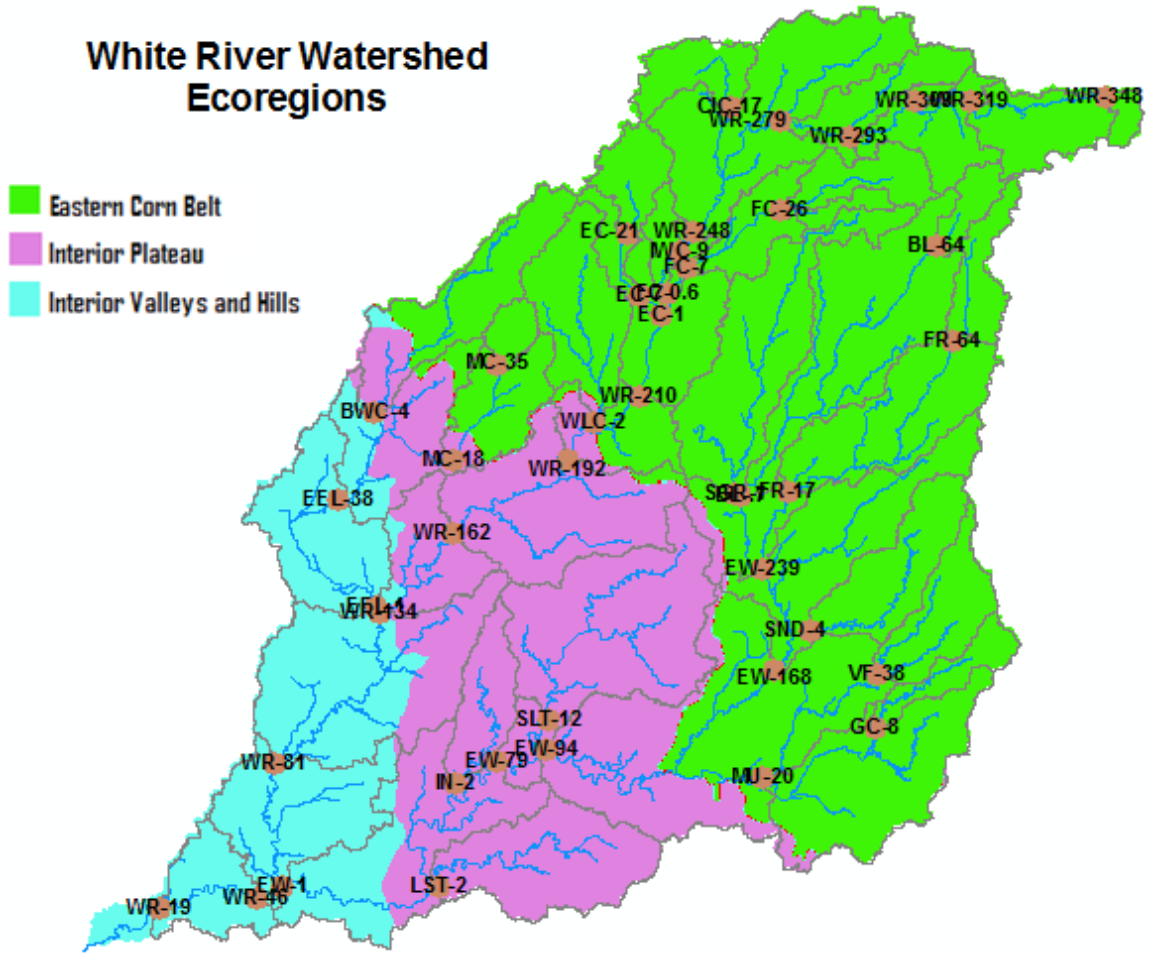


Figure 20 – White River Watershed Level III Ecoregions

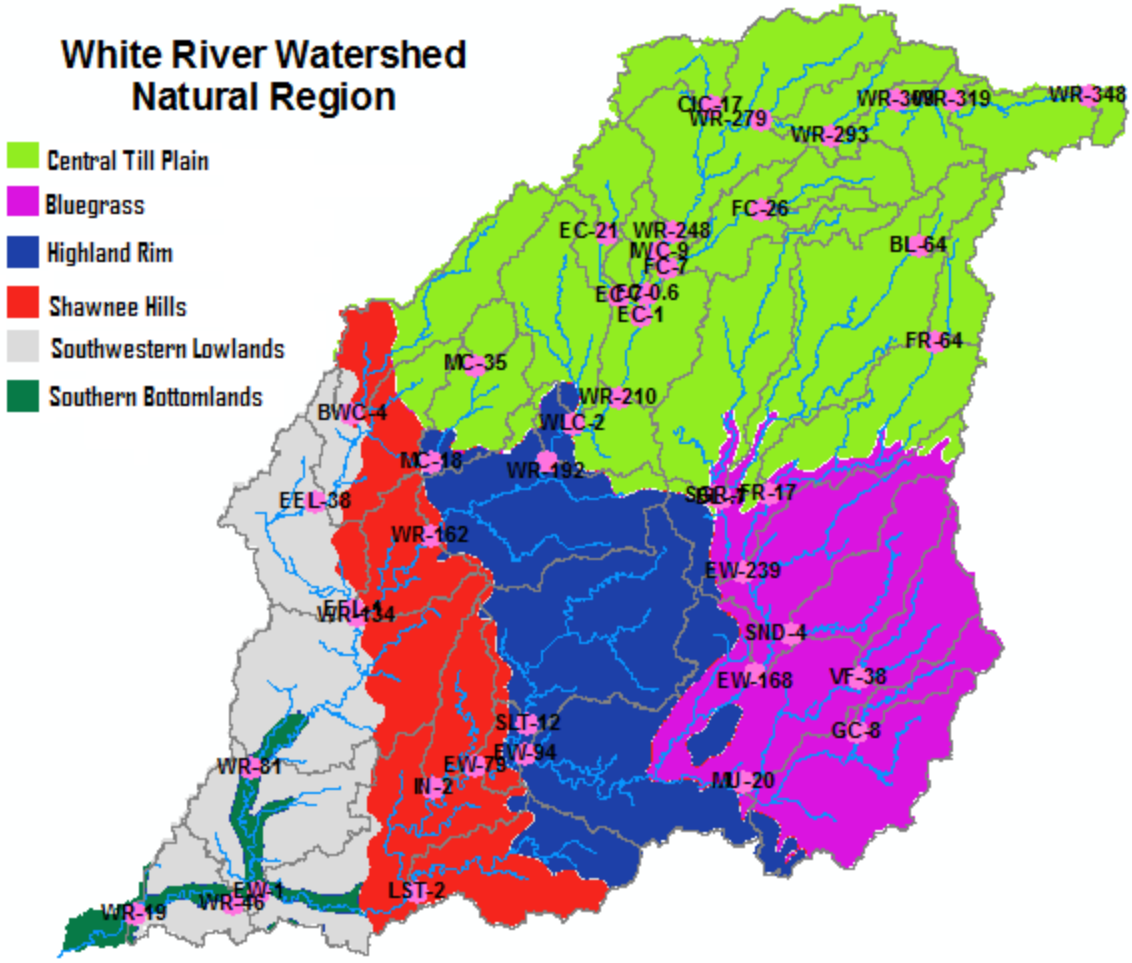


Figure 21 – White River Watershed Natural Regions

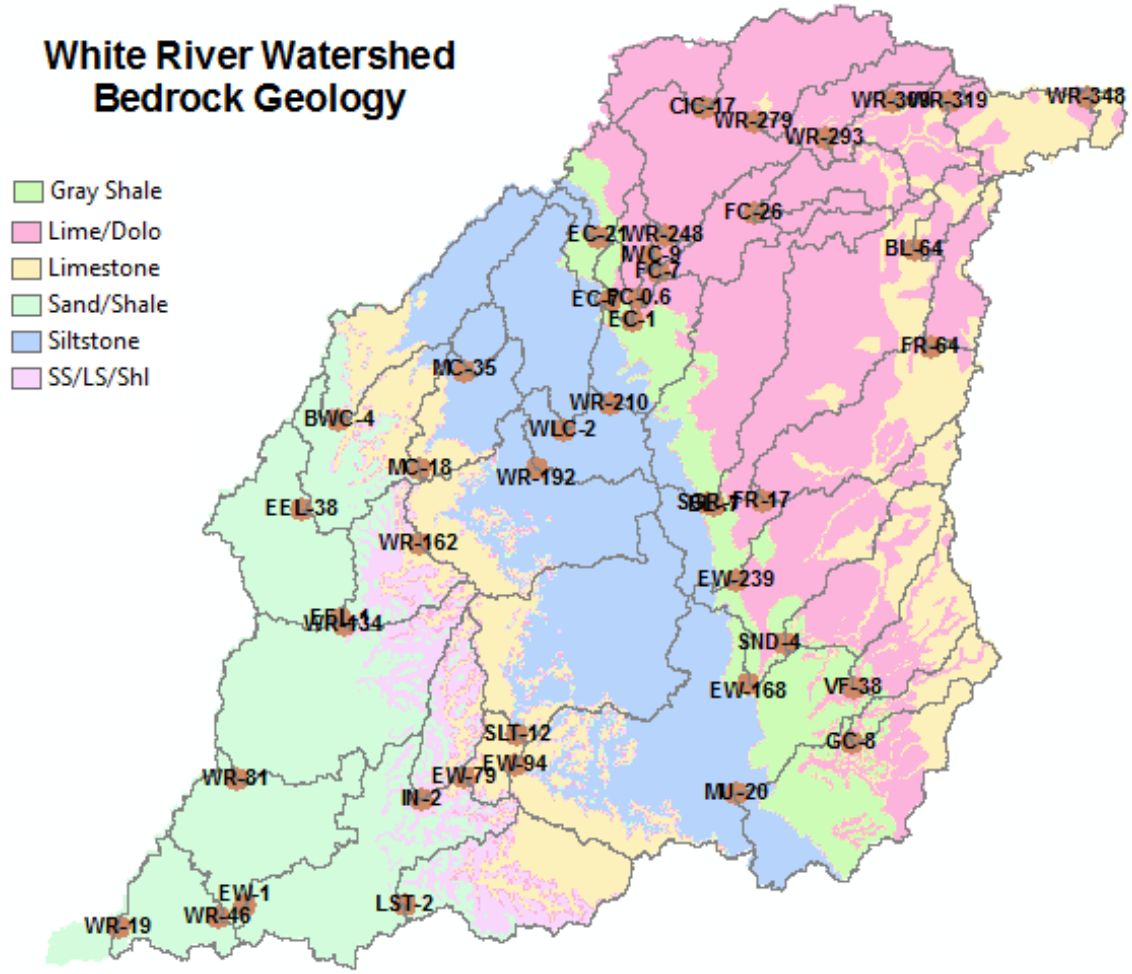


Figure 22 – Bedrock Geology of the White River Watershed; Lime/Dolo=Limestone/Dolomite, Sand/Shale=Sandstone/Shale, SS/LS/Shl=Sandstone/Limestone/Shale

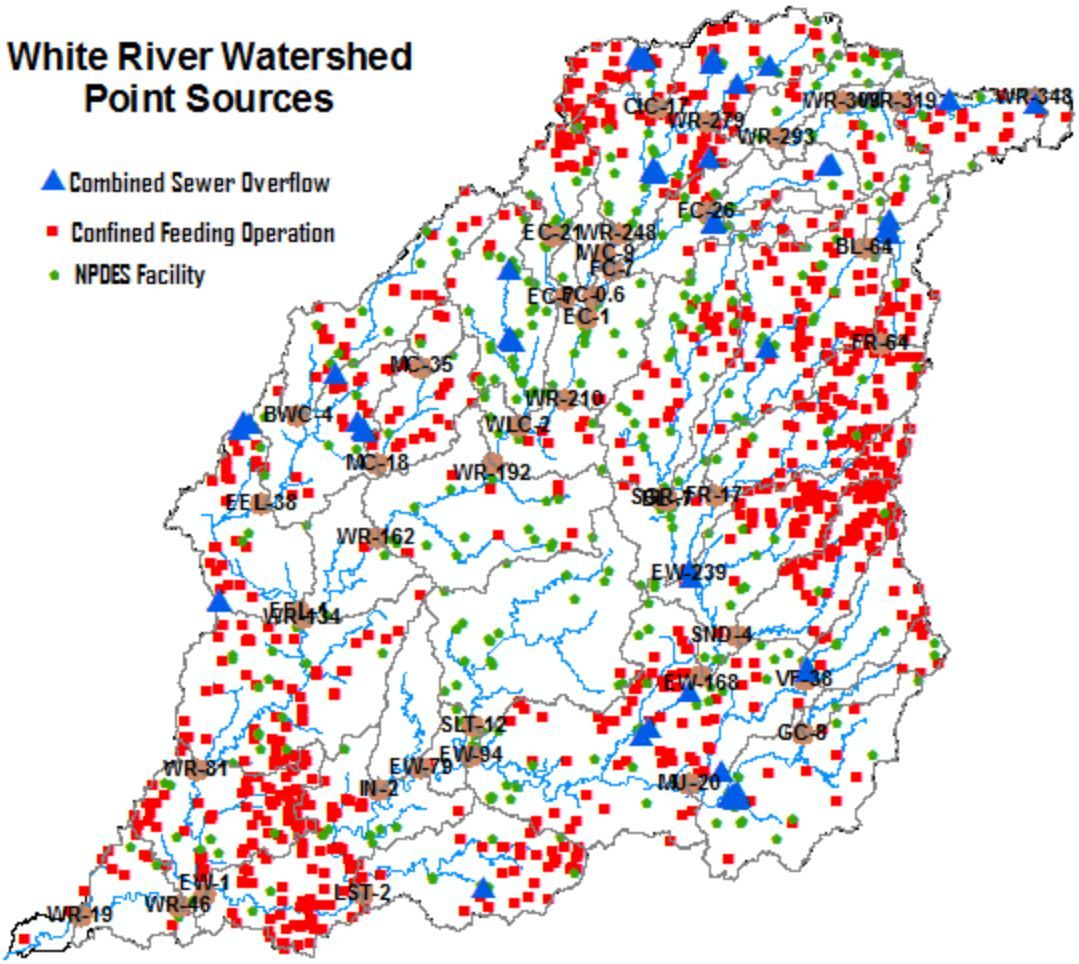


Figure 23 – White River Point Source Pollution; Combined Sewer Overflows, confined feeding operations, NPDES facilities

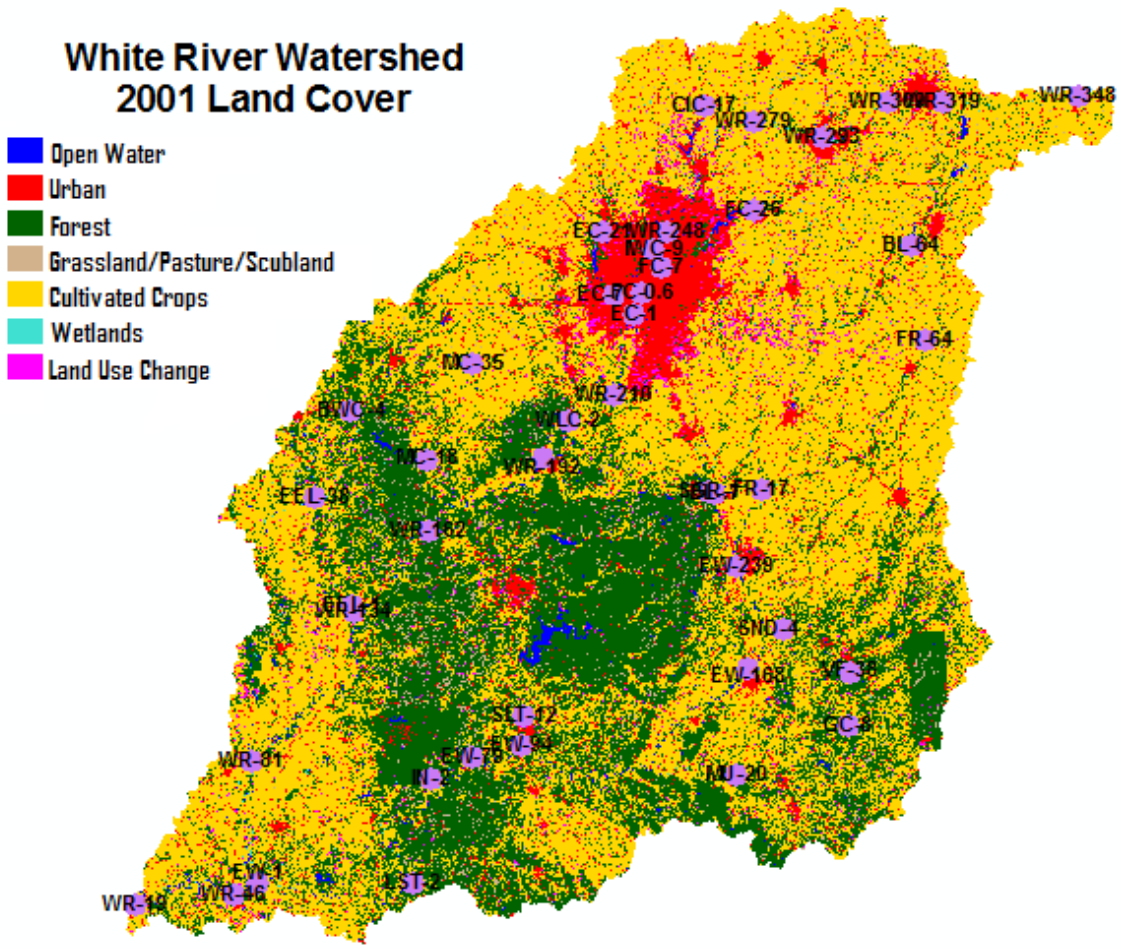


Figure 24 – White River Watershed 2001 land cover dataset

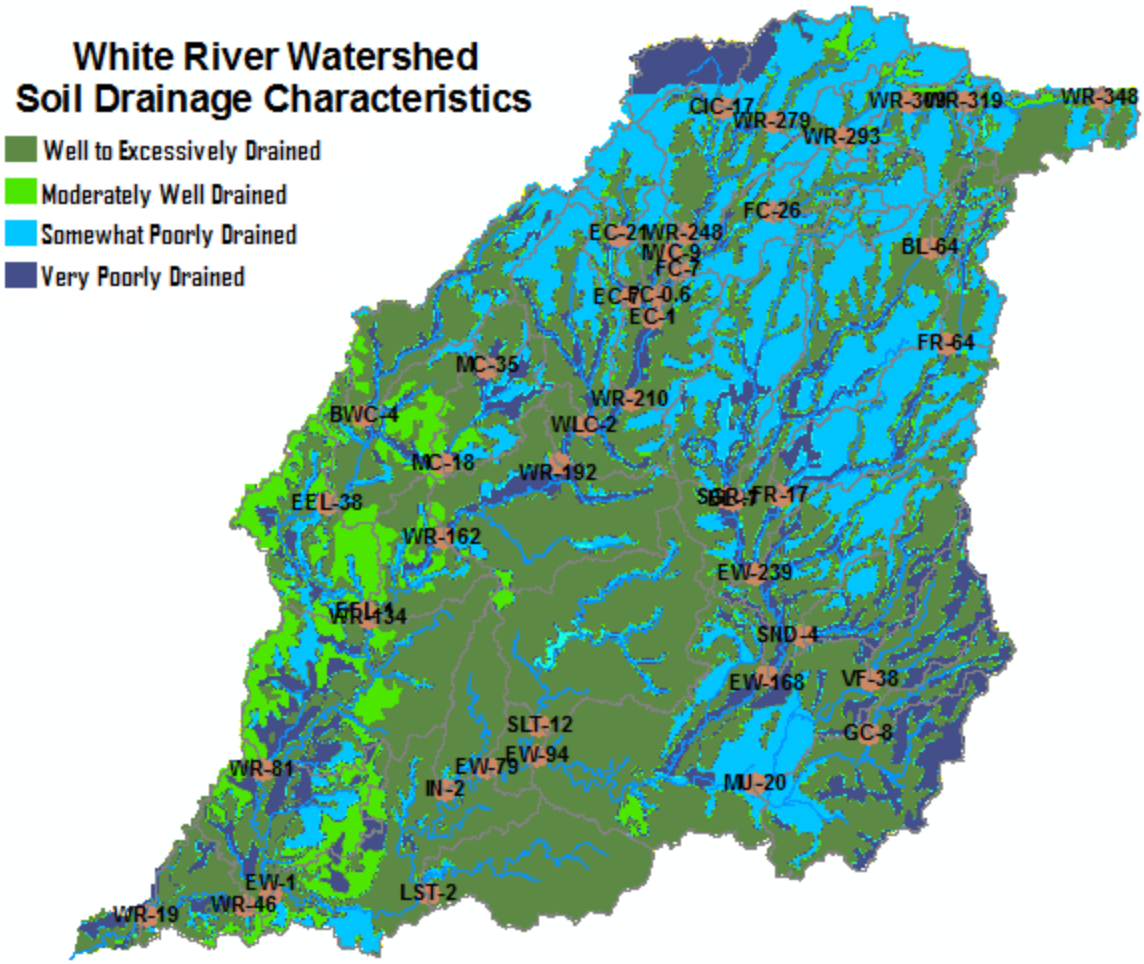


Figure 25 – White River Watershed Soil Drainage Characteristics

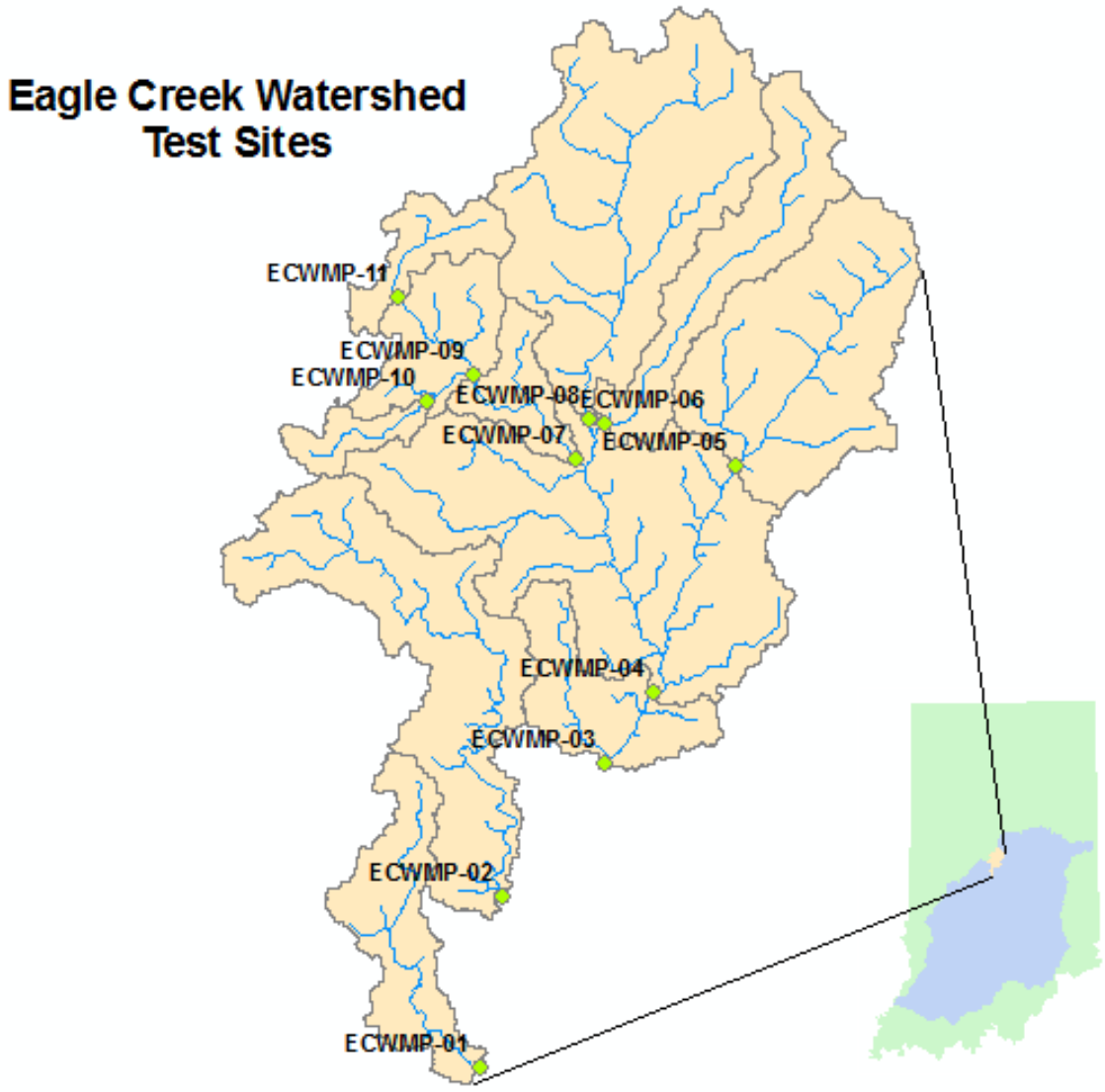


Figure 26 – The Eagle Creek Watershed Monitoring Program (ECWMP) sites for testing model performance

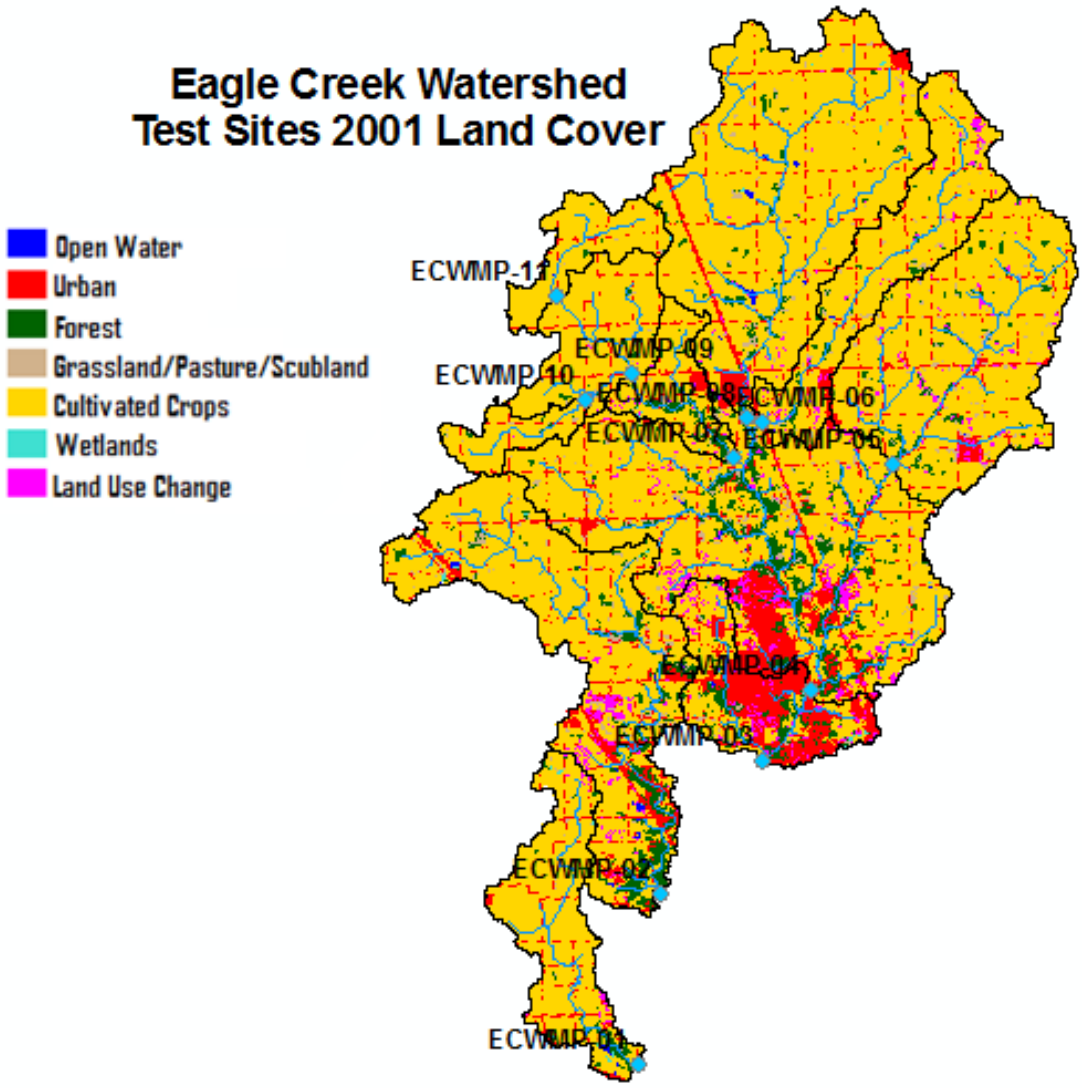


Figure 27 – 2001 Land Use in the ECWMP watersheds

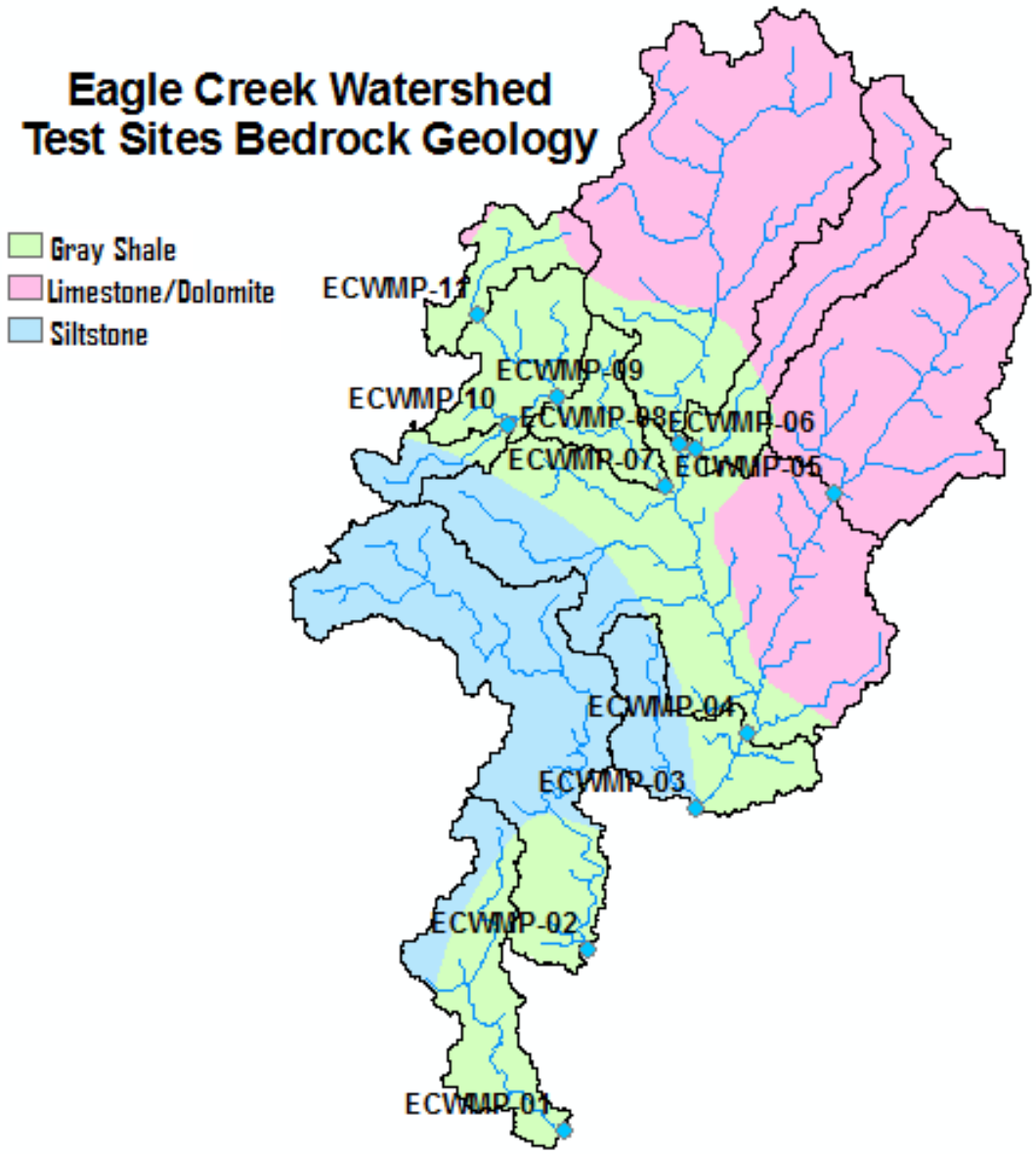


Figure 28 – Bedrock geology in the ECWMP watersheds

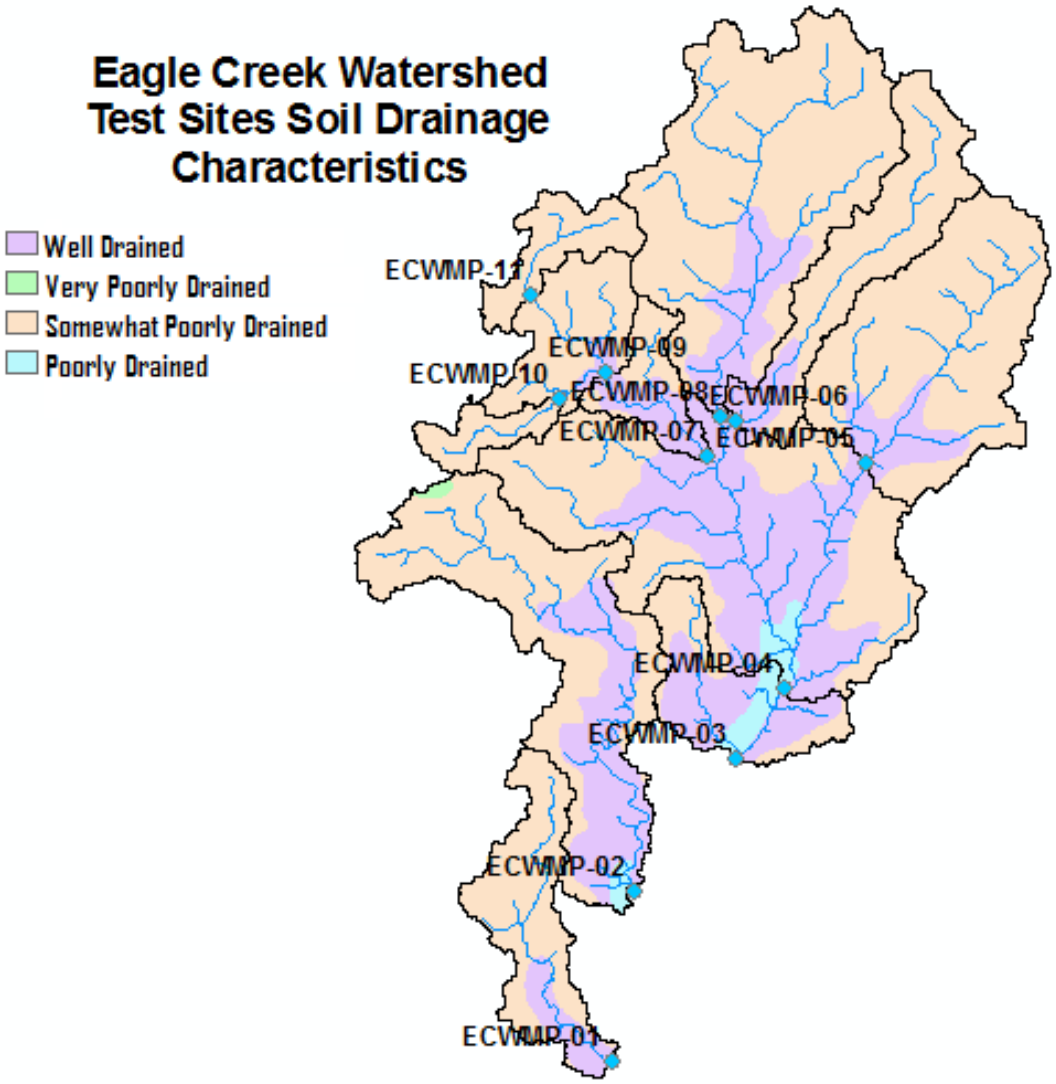


Figure 29 – Soil drainage characteristics of the ECWMP watersheds

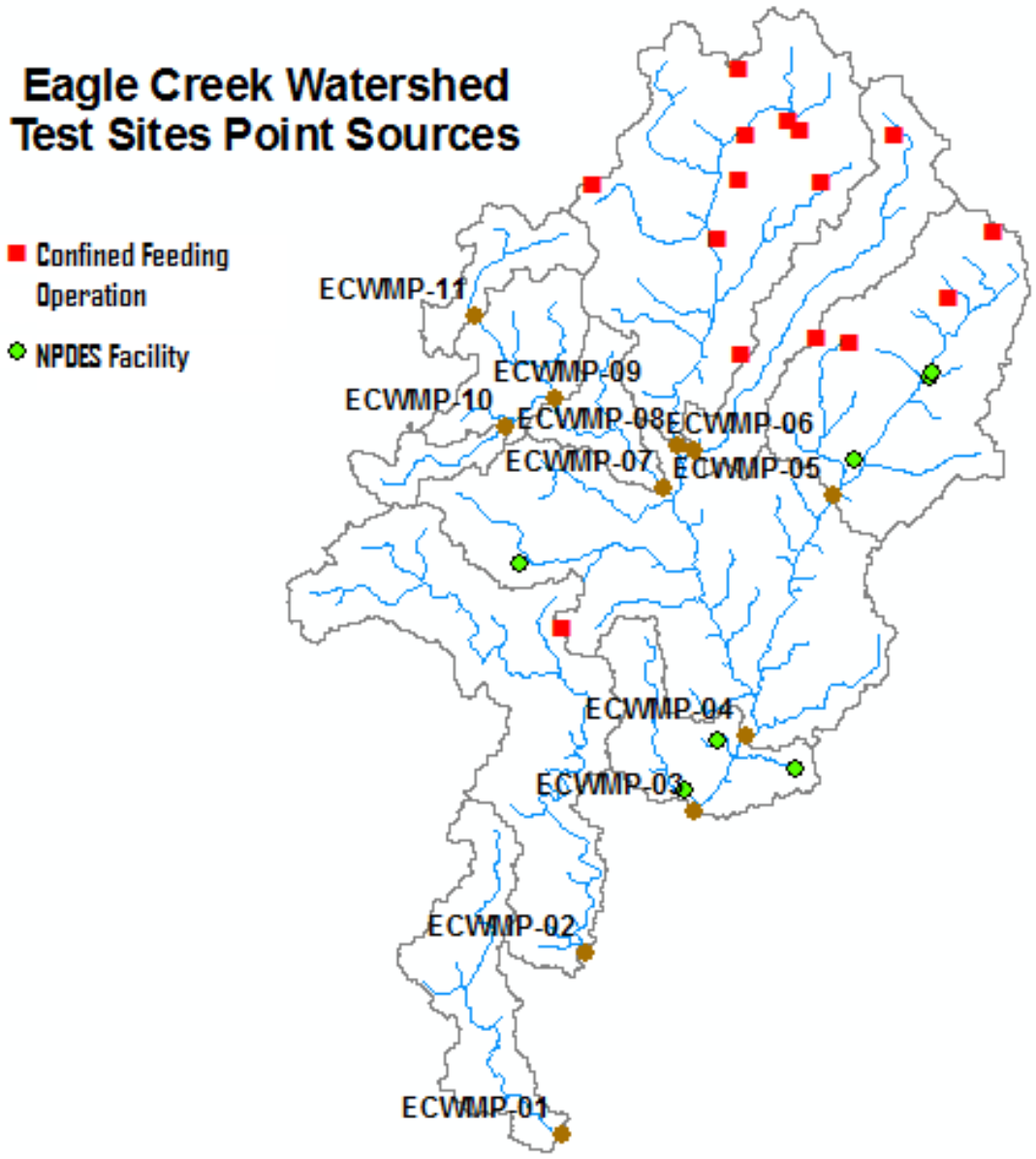


Figure 30 – Point sources in the ECWMP watersheds

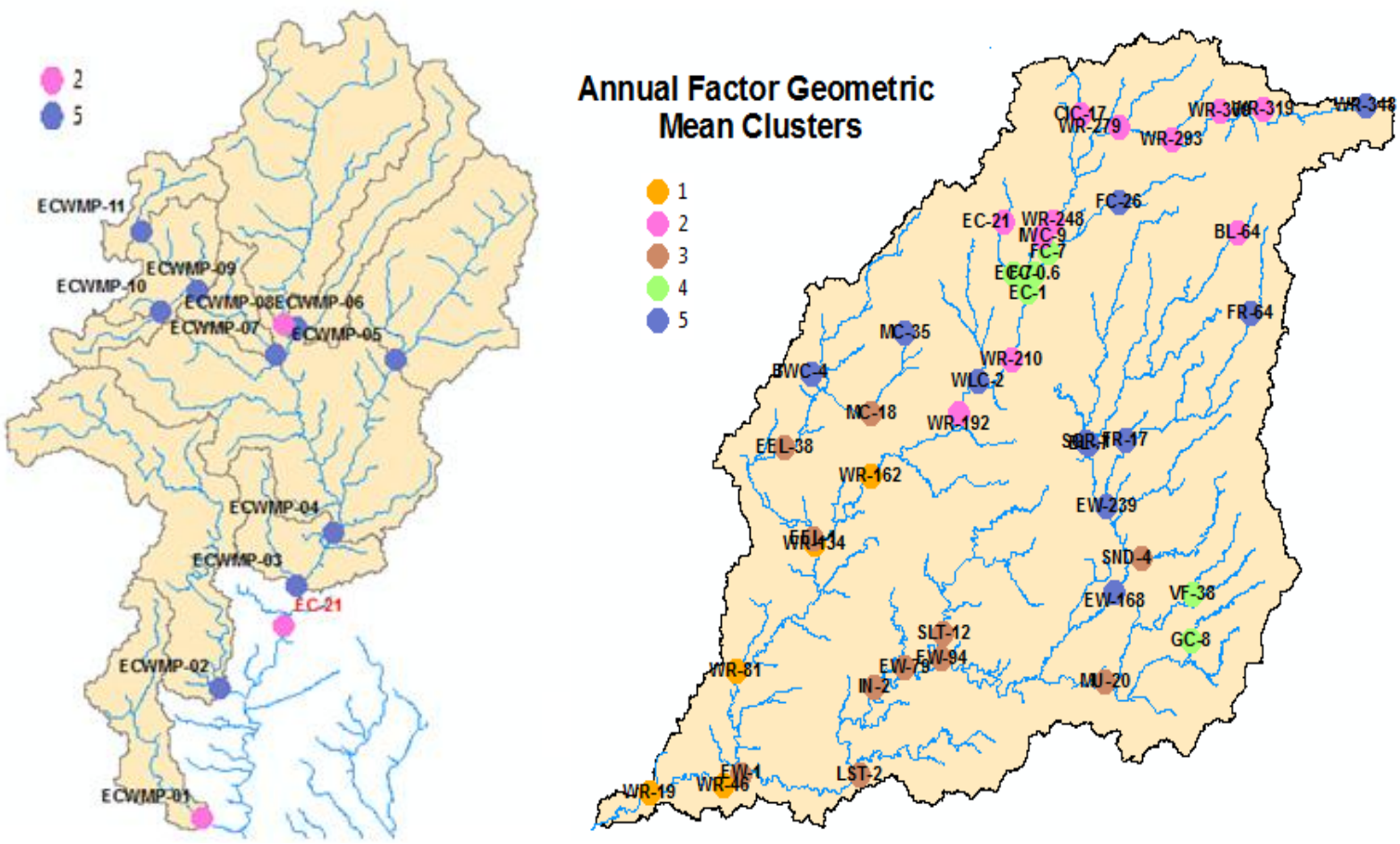


Figure 31 – The LDA annual geometric mean classification model classified the ECWA stations into clusters 2 and 5; the IDEM stations clusters 2 and 5 are located mostly in the upper half of the White River; of note is that ECWA stations, ECWMP-03 and ECWMP-04 were classified into a different cluster than the IDEM station EC-21. These three stations have very similar in watershed characteristics

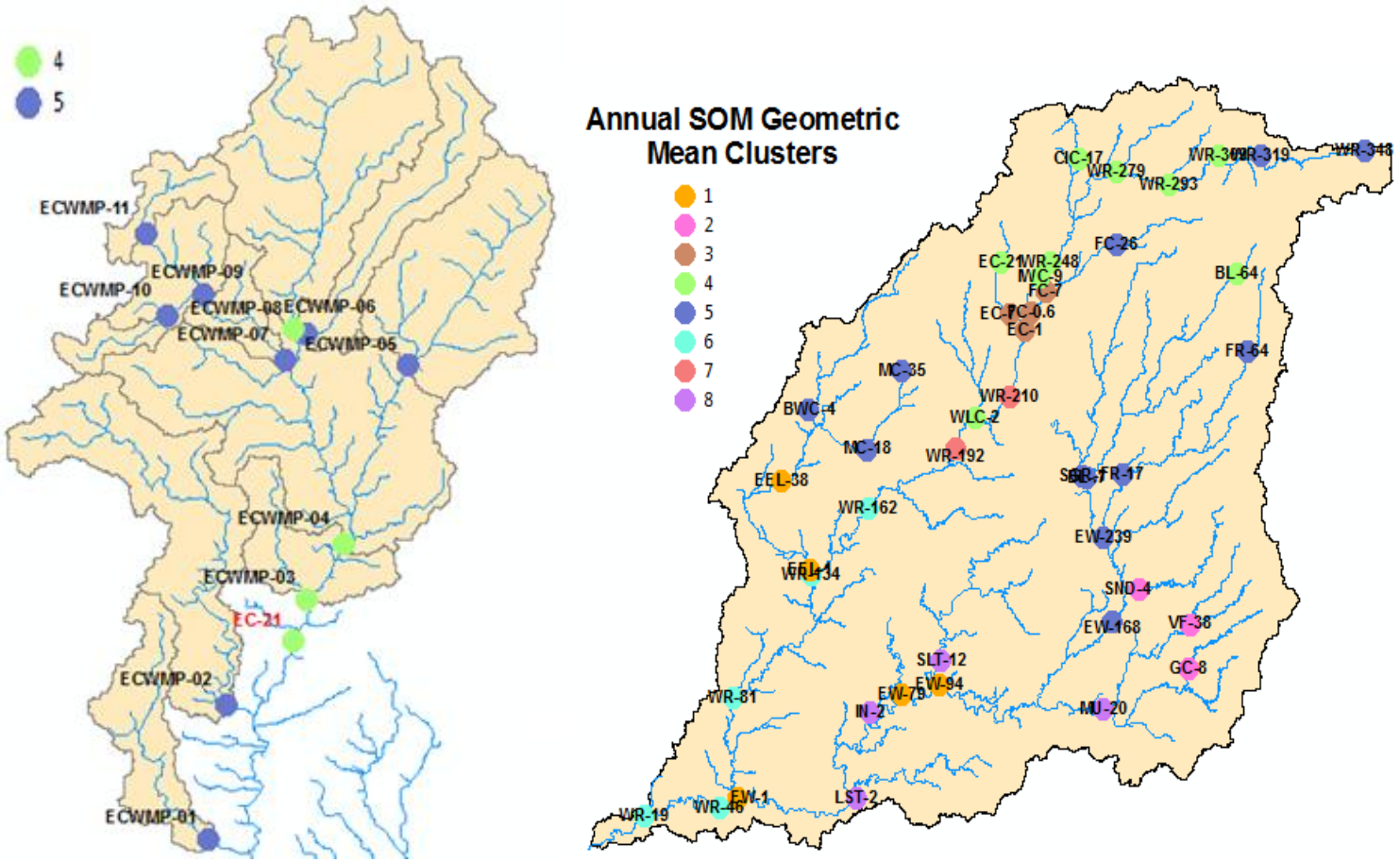


Figure 32 – The SVM annual geometric mean classification model classified the ECWA stations into clusters 4 and 5; the IDEM stations in clusters 4 and 5 are generally located in the upper half of the White River; of note is that ECWA stations, ECWMP-03 and ECWMP-04, were classified into the same cluster as IDEM station, EC-21; these three stations have very similar in watershed characteristics

APPENDIX A – SUPPLEMENTARY TABLES

Box-Cox Transformations

Supplementary Table 1.1 – Annual Box-Cox Transformation Powers and Shapiro-Wilk normality test results; the Shapiro-Wilk normality test was run before and after the Box-Cox transformation was applied

	ANNUAL MEAN			ANNUAL MEDIAN			ANNUAL TRIMMED MEAN			ANNUAL GEOMEAN		
	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest
Alkalinity	0.02	1.90	0.13	0.03	1.65	0.16	0.02	1.85	0.14	0.02	1.90	0.14
TOC	0.21	NA	0.21	0.29	NA	0.29	0.24	NA	0.24	0.17	NA	0.17
Chloride	0.01	0.20	0.71	0.04	0.35	0.76	0.02	0.20	0.72	0.04	0.25	0.72
COD	0.05	NA	0.05	0.05	NA	0.05	0.06	NA	0.06	0.05	0.20	0.11
DO	0.00	5.30	0.25	0.00	4.35	0.13	0.00	5.45	0.16	0.00	4.70	0.09
Hardness	0.02	2.45	0.34	0.02	2.30	0.37	0.02	2.40	0.35	0.02	2.45	0.29
TKN	0.01	-0.40	0.39	0.00	-0.65	0.05	0.01	-0.45	0.38	0.02	-0.40	0.45
NO2 + NO3	0.00	0.20	0.63	0.00	0.30	0.34	0.00	0.25	0.64	0.01	0.20	0.91
pH	0.00	32.30	0.57	0.00	65.40	0.00	0.00	43.65	0.52	0.00	32.10	0.50
Phosphorus	0.00	-0.45	0.65	0.00	-0.25	0.40	0.00	-0.40	0.56	0.00	-0.20	0.46
TSS	0.00	-1.00	0.24	0.00	-0.65	0.30	0.00	-0.90	0.00	0.00	-0.60	0.03
SC	0.75	NA	0.75	0.42	NA	0.42	0.69	NA	0.69	0.55	NA	0.55
SO4	0.00	-0.50	0.74	0.00	-0.55	0.68	0.00	-0.50	0.72	0.00	-0.45	0.73
Temperature	0.69	NA	0.69	0.83	NA	0.83	0.69	NA	0.69	0.62	NA	0.62
Turbidity	0.00	-0.90	0.16	0.00	-1.10	0.01	0.00	-1.20	0.01	0.00	-1.20	0.02
Iron	0.00	-0.95	0.51	0.00	-0.75	0.23	0.00	-0.95	0.01	0.00	-0.85	0.18

Supplementary Table 1.2 – Quarter 1 Box-Cox Transformation Powers and Shapiro-Wilk normality test results; the Shapiro-Wilk normality test was run before and after the Box-Cox transformation was applied

	QUARTER 1 MEAN			QUARTER 1 MEDIAN			QUARTER 1 TRIMMED MEAN			QUARTER 1 GEOMEAN		
	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest
Alkalinity	0.01	1.70	0.04	0.01	1.90	0.07	0.01	1.75	0.03	0.01	1.65	0.05
TOC	0.31	NA	0.31	0.41	NA	0.41	0.32	NA	0.32	0.38	NA	0.38
Chloride	0.01	0.30	0.61	0.10	NA	0.10	0.01	0.30	0.63	0.03	0.35	0.62
COD	0.28	NA	0.28	0.19	NA	0.19	0.38	NA	0.38	0.42	NA	0.42
DO	0.01	2.55	0.07	0.00	8.70	0.69	0.01	3.70	0.14	0.01	1.60	0.02
Hardness	0.00	2.35	0.07	0.00	2.50	0.08	0.00	2.35	0.05	0.00	2.30	0.09
TKN	0.49	NA	0.49	0.37	NA	0.37	0.82	NA	0.82	0.46	NA	0.46
NO2 + NO3	0.02	0.15	0.04	0.05	0.35	0.05	0.01	0.15	0.02	0.03	0.25	0.04
pH	0.01	26.65	1.00	0.04	7.85	0.17	0.01	25.50	0.97	0.01	26.40	1.00
Phosphorus	0.00	-0.20	0.52	0.00	0.10	0.37	0.00	-0.25	0.10	0.00	0.00	0.15
TSS	0.00	-0.55	0.11	0.00	-0.70	0.07	0.00	-0.85	0.06	0.00	-0.60	0.11
SC	0.11	NA	0.11	0.09	NA	0.09	0.07	NA	0.07	0.09	NA	0.09
SO4	0.00	-1.00	0.66	0.00	-1.00	0.63	0.00	-0.85	0.81	0.00	-0.85	0.83
Temperature	0.60	NA	0.60	0.33	NA	0.33	0.77	NA	0.77	0.62	NA	0.62
Turbidity	0.00	-0.90	0.01	0.00	-0.50	0.55	0.00	-1.15	0.00	0.00	-0.85	0.17
Iron	0.00	-0.75	0.11	0.00	0.40	0.01	0.00	0.55	0.04	0.00	0.45	0.00

Supplementary Table 1.3 – Quarter 2 Box-Cox Transformation Powers and Shapiro-Wilk normality test results; the Shapiro-Wilk normality test was run before and after the Box-Cox transformation was applied

	QUARTER 2 MEAN			QUARTER 2 MEDIAN			QUARTER 2 TRIMMED MEAN			QUARTER 2 GEOMEAN		
	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest
Alkalinity	0.01	1.80	0.07	0.01	1.50	0.07	0.01	1.70	0.06	0.01	1.90	0.06
TOC	0.10	NA	0.10	0.59	NA	0.59	0.14	NA	0.14	0.39	NA	0.39
Chloride	0.03	0.30	0.75	0.04	0.35	0.80	0.03	0.30	0.73	0.06	NA	0.06
COD	0.27	NA	0.27	0.19	NA	0.19	0.23	NA	0.23	0.26	NA	0.26
DO	0.00	3.45	0.04	0.00	0.90	0.00	0.00	1.65	0.00	0.00	2.30	0.00
Hardness	0.00	2.45	0.11	0.00	2.20	0.09	0.00	2.40	0.09	0.00	2.45	0.09
TKN	0.33	NA	0.33	0.11	NA	0.11	0.15	NA	0.15	0.17	NA	0.17
NO2 + NO3	0.01	0.20	0.06	0.00	0.20	0.04	0.01	0.25	0.08	0.01	0.35	0.20
pH	0.00	4.00	0.02	0.00	40.85	0.00	0.00	15.60	0.48	0.01	4.15	0.02
Phosphorus	0.01	-0.20	0.66	0.00	-0.05	0.64	0.00	-0.20	0.46	0.00	-0.10	0.44
TSS	0.00	-0.10	0.63	0.00	-0.45	0.14	0.00	-0.25	0.35	0.00	-0.50	0.01
SC	0.30	NA	0.30	0.13	NA	0.13	0.19	NA	0.19	0.09	NA	0.09
SO4	0.00	-0.40	0.98	0.00	-0.40	0.96	0.00	-0.40	0.96	0.00	-0.40	0.97
Temperature	0.49	NA	0.49	0.47	NA	0.47	0.54	NA	0.54	0.48	NA	0.48
Turbidity	0.01	0.15	0.44	0.00	-0.70	0.28	0.00	-0.10	0.33	0.00	-0.70	0.26
Iron	0.00	0.05	0.55	0.00	-0.35	0.71	0.00	0.00	0.54	0.00	-0.40	0.83

Supplementary Table 1.4 – Quarter 3 Box-Cox Transformation Powers and Shapiro-Wilk normality test results; the Shapiro-Wilk normality test was run before and after the Box-Cox transformation was applied

	QUARTER 3 MEAN			QUARTER 3 MEDIAN			QUARTER 3 TRIMMED MEAN			QUARTER 3 GEOMEAN		
	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest
Alkalinity	0.09	NA	0.09	0.08	NA	0.08	0.09	NA	0.09	0.11	NA	0.11
TOC	0.11	NA	0.11	0.22	NA	0.22	0.16	NA	0.16	0.12	NA	0.12
Chloride	0.00	0.05	0.38	0.00	0.15	0.29	0.00	0.05	0.38	0.00	0.10	0.45
COD	0.05	NA	0.05	0.01	0.20	0.14	0.03	-0.25	0.51	0.03	-0.15	0.40
DO	0.01	1.95	0.07	0.00	2.60	0.14	0.00	2.10	0.06	0.00	2.20	0.05
Hardness	0.12	NA	0.12	0.06	NA	0.06	0.12	NA	0.12	0.15	2.55	0.32
TKN	0.00	-0.60	0.28	0.00	-0.45	0.03	0.00	-0.55	0.31	0.00	-0.50	0.30
NO2 + NO3	0.00	0.30	0.12	0.00	0.50	0.06	0.00	0.30	0.12	0.00	0.30	0.16
pH	0.00	5.80	0.01	0.00	5.10	0.00	0.00	9.60	0.03	0.00	6.20	0.01
Phosphorus	0.00	-0.40	0.50	0.00	-0.30	0.34	0.00	-0.40	0.51	0.00	-0.35	0.56
TSS	0.00	-0.35	0.04	0.00	-0.65	0.07	0.00	-0.50	0.00	0.00	-0.65	0.00
SC	0.55	NA	0.55	0.52	NA	0.52	0.60	NA	0.60	0.79	NA	0.79
SO4	0.00	-0.25	0.52	0.00	-0.25	0.65	0.00	-0.25	0.55	0.00	-0.20	0.59
Temperature	0.85	NA	0.85	1.00	NA	1.00	0.80	NA	0.80	0.85	NA	0.85
Turbidity	0.00	-0.30	0.13	0.00	-1.00	0.00	0.00	-0.65	0.01	0.00	-0.70	0.01
Iron	0.00	-0.30	0.04	0.00	-0.75	0.01	0.00	-0.50	0.01	0.00	-0.55	0.02

Supplementary Table 1.5 – Quarter 4 Box-Cox Transformation Powers and Shapiro-Wilk normality test results; the Shapiro-Wilk normality test was run before and after the Box-Cox transformation was applied

	QUARTER 4 MEAN			QUARTER 4 MEDIAN			QUARTER 4 TRIMMED MEAN			QUARTER 4 GEOMEAN		
	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest	SW Pretest	Box-Cox	SW Postest
Alkalinity	0.03	1.60	0.07	0.02	1.85	0.10	0.03	1.60	0.09	0.02	1.60	0.07
TOC	0.37	NA	0.37	0.44	NA	0.44	0.29	NA	0.29	0.32	NA	0.32
Chloride	0.01	0.25	0.76	0.01	0.35	0.41	0.02	0.25	0.80	0.05	0.30	0.90
COD	0.14	NA	0.14	0.03	0.30	0.05	0.15	NA	0.15	0.05	NA	0.05
DO	0.00	8.10	0.36	0.00	6.50	0.12	0.00	8.35	0.43	0.00	8.30	0.33
Hardness	0.07	NA	0.07	0.10	NA	0.10	0.07	NA	0.07	0.05	NA	0.05
TKN	0.00	-0.55	0.36	0.00	-0.60	0.08	0.00	-0.45	0.29	0.00	-0.55	0.36
NO2 + NO3	0.00	0.25	0.50	0.00	0.45	0.59	0.00	0.25	0.51	0.01	0.15	0.49
pH	0.03	32.35	0.08	0.00	28.75	0.13	0.01	37.10	0.02	0.03	32.15	0.09
Phosphorus	0.00	-0.35	0.34	0.00	-0.15	0.09	0.00	-0.35	0.28	0.00	-0.10	0.15
TSS	0.00	-0.55	0.63	0.00	0.05	0.08	0.00	-0.55	0.61	0.00	0.00	0.58
SC	0.65	NA	0.65	0.59	NA	0.59	0.69	NA	0.69	0.68	NA	0.68
SO4	0.00	-0.50	0.47	0.00	-0.55	0.49	0.00	-0.55	0.38	0.00	-0.50	0.37
Temperature	0.36	NA	0.36	0.20	NA	0.20	0.31	NA	0.31	0.35	NA	0.35
Turbidity	0.00	-0.50	0.11	0.00	-0.65	0.41	0.00	-0.65	0.28	0.00	-0.75	0.33
Iron	0.00	-0.60	0.82	0.00	-0.20	0.50	0.00	-0.50	0.50	0.00	-0.55	0.23

PCA Loadings

Supplementary Table 2.1 – Annual Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**90**	-22	-20	16	
TOC	-16	**89**	-2	-2	
Chloride	**75**	54	-17	16	
COD	-12	**92**	32	7	
DO	10	24	30	**85**	
Hardness	**94**	-13	-21	12	
TKN	24	**89**	26	13	
NO2 + NO3	**83**	-19	-5	9	
pH	17	-8	30	**87**	
Total P	54	**63**	42	3	
TSS	-9	22	**93**	-4	
SC	**92**	32	-14	9	
Sulfate	**73**	47	10	-17	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-19	9	**93**	13	
Iron	-14	16	**96**	-9	Final Communality
Eigenvalue	4.802	3.702	3.307	1.645	13.456795
% Variance Explained	32.0133	24.68	22.0467	10.9667	89.7119667

Supplementary Table 2.2 – Annual Median Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**87**	-31	-26	16	
TOC	-6	2	**93**	-7	
Chloride	**81**	-3	46	22	
COD	-6	43	**87**	9	
DO	9	-16	39	**78**	
Hardness	**92**	-29	-15	12	
TKN	25	39	**83**	12	
NO2 + NO3	**77**	-15	-24	3	
pH	22	5	-19	**86**	
Total P	DNL	DNL	DNL	DNL	
TSS	-8	**95**	16	3	
SC	**95**	-14	21	14	
Sulfate	**82**	16	35	-1	
Temperature	-9	**79**	19	1	
Turbidity	-18	**94**	12	-3	
Iron	-19	**92**	10	-16	Final Communality
Eigenvalue	4.624	3.879	3.127	1.515	13.146
% Variance Explained	30.82667	25.86	20.84667	10.1	87.64

Supplementary Table 2.3 – Annual Trimmed Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**88**	-26	-28	15	
TOC	-9	1	**91**	-5	
Chloride	**80**	-9	48	18	
COD	-5	39	**90**	7	
DO	10	-32	27	**83**	
Hardness	**93**	-25	-19	11	
TKN	29	33	**85**	12	
NO2 + NO3	**81**	-8	-25	7	
pH	16	21	-11	**89**	
Total P	DNL	DNL	DNL	DNL	
TSS	-4	**97**	16	-3	
SC	**94**	-12	25	11	
Sulfate	**78**	14	39	-11	
Temperature	-21	**66**	47	12	
Turbidity	-15	**95**	12	2	
Iron	-16	**95**	10	-11	Final Communality
Eigenvalue	4.657	3.758	3.344	1.63	13.389
% Variance Explained	31.0467	23.4875	20.9	10.1875	83.68125

Supplementary Table 2.4 – Annual Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**86**	-32	-25	14	
TOC	-6	3	**95**	-5	
Chloride	**82**	-4	45	20	
COD	-3	39	**90**	7	
DO	16	-23	26	**82**	
Hardness	**91**	-30	-17	11	
TKN	30	37	**84**	12	
NO2 + NO3	**83**	-10	-19	7	
pH	12	10	-12	**89**	
Total P	DNL	DNL	DNL	DNL	
TSS	-4	**97**	15	-2	
SC	**95**	-11	22	14	
Sulfate	**81**	19	35	-9	
Temperature	-18	**75**	40	16	
Turbidity	-16	**94**	11	-4	
Iron	-14	**94**	10	-21	Final Communality
Eigenvalue	4.712	3.878	3.208	1.663	13.46
% Variance Explained	31.4133	25.8533	21.3867	11.0867	89.73333333

Supplementary Table 2.5 – Quarter 1 Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**92**	-11	-25	16	
TOC	-14	**95**	7	-7	
Chloride	**78**	52	-12	-4	
COD	-3	**91**	35	-11	
DO	-8	-9	-21	**84**	
Hardness	**94**	-9	-23	17	
TKN	39	**82**	29	-13	
NO2 + NO3	**74**	-33	5	22	
pH	31	-9	-11	**78**	
Total P	42	**66**	52	-6	
TSS	-8	18	**91**	-21	
SC	**91**	29	-8	2	
Sulfate	**80**	39	7	-22	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-19	19	**94**	-5	
Iron	-16	17	**93**	-17	Final Communality
Eigenvalue	4.861	3.596	3.257	1.577	13.291
% Variance Explained	32.4067	23.9733	21.71333	10.51333	88.60666667

Supplementary Table 2.6 – Quarter 1 Median Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	
Alkalinity	**89**	-18	-27	19	-9	
TOC	-5	**97**	7	-8	3	
Chloride	**79**	50	-13	-1	0	
COD	-3	**91**	33	-7	11	
DO	-13	-23	-37	**69**	-33	
Hardness	**91**	-18	-27	16	-11	
TKN	42	**77**	32	-9	11	
NO2 + NO3	**67**	-45	-8	17	3	
pH	38	-2	-22	**82**	4	
Total P	DNL	DNL	DNL	DNL	DNL	
TSS	0	13	**90**	-24	15	
SC	**95**	16	-8	10	-1	
Sulfate	**85**	35	7	-14	-2	
Temperature	-11	10	20	-10	**95**	
Turbidity	-23	34	**83**	-18	16	
Iron	-34	12	**87**	-11	3	Final Communality
Eigenvalue	4.692	3.199	2.94	1.366	1.244	13.442
% Variance Explained	31.28	21.3267	19.6	9.10667	8.29333	89.61333333

Supplementary Table 2.7 – Quarter 1 Trimmed Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**91**	-13	-27	18	
TOC	-12	**96**	3	-8	
Chloride	**79**	50	-16	-2	
COD	-5	**91**	34	-16	
DO	-9	-10	-15	**86**	
Hardness	**93**	-10	-25	19	
TKN	39	**82**	29	-18	
NO2 + NO3	**74**	-35	8	23	
pH	34	-10	-8	**69**	
Total P	43	**65**	52	-11	
TSS	-11	12	**90**	-31	
SC	**92**	25	-12	0	
Sulfate	**80**	37	0	-24	
Temperature	-3	12	32	**-65**	
Turbidity	-17	30	**88**	-13	
Iron	-27	10	**85**	-16	Final Communality
Eigenvalue	4.947	3.589	3.1	2.017	13.654
% Variance Explained	30.91875	22.4313	19.375	12.6063	85.3375

Supplementary Table 2.8 – Quarter 1 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**91**	-11	-31	11	
TOC	-10	**96**	6	-9	
Chloride	**79**	51	-11	-1	
COD	-3	**92**	32	-11	
DO	-11	-10	-25	**84**	
Hardness	**92**	-10	-30	14	
TKN	37	**84**	29	-12	
NO2 + NO3	**73**	-35	2	25	
pH	30	-10	-22	**71**	
Total P	42	**63**	57	-4	
TSS	-8	18	**90**	-29	
SC	**91**	27	-9	3	
Sulfate	**79**	40	5	-23	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-22	29	**87**	-18	
Iron	-32	8	**85**	-17	Final Communality
Eigenvalue	4.898	3.653	3.116	1.541	13.208
% Variance Explained	32.65333	24.3533	20.77333	10.2733	88.05333333

Supplementary Table 2.9 – Quarter 2 Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**88**	-29	-15	19	
TOC	-7	**83**	-2	5	
Chloride	**85**	39	-19	17	
COD	3	**91**	32	-1	
DO	32	26	-25	**73**	
Hardness	**92**	-25	-16	15	
TKN	39	**81**	38	3	
NO2 + NO3	**70**	-44	16	22	
pH	9	-5	31	**84**	
Total P	DNL	DNL	DNL	DNL	
TSS	-6	26	**94**	0	
SC	**97**	12	-13	13	
Sulfate	**82**	35	-2	-2	
Temperature	-18	**76**	31	12	
Turbidity	-23	14	**91**	18	
Iron	-9	20	**96**	-4	Final Communality
Eigenvalue	4.802	3.584	3.261	1.513	13.161
% Variance Explained	32.01333	23.89333	21.74	10.08667	87.74

Supplementary Table 2.10 – Quarter 2 Median Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**89**	-26	-22	19	
TOC	-6	1	**92**	-4	
Chloride	**81**	-16	46	20	
COD	-6	39	**90**	6	
DO	27	-46	26	**63**	
Hardness	**93**	-22	-17	16	
TKN	27	39	**85**	4	
NO2 + NO3	**69**	9	-39	29	
pH	19	5	-8	**88**	
Total P	DNL	DNL	DNL	DNL	
TSS	-9	**93**	24	-1	
SC	**95**	-17	17	16	
Sulfate	**81**	3	41	-13	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-22	**94**	12	2	
Iron	-12	**94**	18	-15	Final Communality
Eigenvalue	4.589	3.344	3.186	1.447	12.567
% Variance Explained	32.77857	23.88571	22.75714	10.33571	89.76428571

Supplementary Table 2.11 – Quarter 2 Trimmed Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**89**	-29	-18	17	
TOC	-5	**87**	1	-4	
Chloride	**85**	39	-18	18	
COD	1	**91**	35	2	
DO	33	21	-33	**70**	
Hardness	**92**	-25	-18	13	
TKN	35	**83**	37	3	
NO2 + NO3	**70**	-44	17	20	
pH	8	-4	21	**89**	
Total P	DNL	DNL	DNL	DNL	
TSS	-7	24	**95**	-1	
SC	**97**	11	-13	12	
Sulfate	**81**	36	-5	-19	
Temperature	-21	**75**	31	21	
Turbidity	-22	20	**93**	10	
Iron	-11	20	96	-5	Final Communality
Eigenvalue	4.793	3.645	3.356	1.507	13.273
% Variance Explained	31.95333	24.3	22.37333	10.04667	88.48666667

Supplementary Table 2.12 – Quarter 2 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**88**	-25	-24	16	
TOC	-6	**89**	-3	-7	
Chloride	**80**	46	-24	8	
COD	1	**91**	34	1	
DO	34	26	-40	**63**	
Hardness	**92**	-22	-23	13	
TKN	33	**85**	36	6	
NO2 + NO3	**71**	-43	11	24	
pH	7	-3	6	**92**	
Total P	DNL	DNL	DNL	DNL	
TSS	-8	26	**93**	-3	
SC	**91**	12	-16	12	
Sulfate	**82**	36	4	-20	
Temperature	-20	**75**	37	22	
Turbidity	-23	16	**93**	5	
Iron	-12	19	**95**	-13	Final Communality
Eigenvalue	4.716	3.74	3.376	1.476	13.309
% Variance Explained	31.44	24.93333	22.50667	9.84	88.72666667

Supplementary Table 2.13 – Quarter 3 Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**82**	-25	-39	14	
TOC	-1	-6	**93**	-5	
Chloride	**81**	-5	43	26	
COD	-1	41	**86**	24	
DO	13	2	23	**90**	
Hardness	**91**	-20	-25	12	
TKN	22	35	**82**	29	
NO2 + NO3	**84**	-19	-11	-5	
pH	12	17	1	**91**	
Total P	**71**	34	48	12	
TSS	-5	**95**	23	9	
SC	**94**	-9	26	13	
Sulfate	**79**	24	36	0	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-10	**92**	6	14	
Iron	-13	**96**	12	-1	Final Communality
Eigenvalue	4.982	3.317	3.244	1.93	13.473
% Variance Explained	33.21333	22.1133	21.62667	12.8667	89.82

Supplementary Table 2.14 – Quarter 3 Median Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**75**	-30	-48	18	
TOC	8	-4	**94**	-8	
Chloride	**85**	-3	35	22	
COD	3	39	**88**	20	
DO	14	-9	24	**90**	
Hardness	**88**	-23	-30	17	
TKN	28	41	**80**	24	
NO2 + NO3	**81**	-16	-17	-7	
pH	12	15	-5	**89**	
Total P	**76**	29	38	7	
TSS	3	**97**	17	11	
SC	**96**	-13	15	11	
Sulfate	**84**	16	26	-1	
Temperature	-25	**61**	53	28	
Turbidity	-10	**95**	11	-2	
Iron	-8	**97**	13	-6	Final Communality
Eigenvalue	5.142	3.79	3.401	1.921	14.25
% Variance Explained	32.137	23.687	21.2562	12.006	89.0625

Supplementary Table 2.15 – Quarter 3 Trimmed Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**81**	-25	-41	13	
TOC	1	-9	**95**	-4	
Chloride	**82**	-3	40	27	
COD	-3	36	**89**	21	
DO	13	-2	22	**91**	
Hardness	**91**	-19	-27	12	
TKN	23	34	**83**	30	
NO2 + NO3	**84**	-20	-12	-4	
pH	13	17	3	**90**	
Total P	**72**	31	48	11	
TSS	-2	**96**	21	11	
SC	**95**	-8	23	13	
Sulfate	**80**	23	35	1	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-9	**95**	8	10	
Iron	-13	**97**	9	-2	Final Communality
Eigenvalue	5.035	3.348	3.308	1.927	13.618
% Variance Explained	33.56667	22.32	22.05333	12.84667	90.78666667

Supplementary Table 2.16 – Quarter 3 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**81**	-26	-39	13	
TOC	1	-8	**95**	-5	
Chloride	**83**	0	38	27	
COD	-3	38	**90**	18	
DO	14	1	19	**90**	
Hardness	**90**	-25	-26	5	
TKN	23	37	**83**	28	
NO2 + NO3	**84**	-14	-12	-1	
pH	12	13	1	**91**	
Total P	**71**	30	48	11	
TSS	-2	**96**	20	15	
SC	**94**	-7	21	16	
Sulfate	**82**	21	34	2	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-11	**96**	10	8	
Iron	-11	**97**	11	-5	Final Communality
Eigenvalue	5.044	3.368	3.274	1.911	13.597
% Variance Explained	33.6267	22.4533	21.8267	12.74	90.64666667

Supplementary Table 2.17 – Quarter 4 Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**78**	-32	-36	21	
TOC	-8	12	**88**	-13	
Chloride	**80**	-21	43	8	
COD	-3	40	**89**	1	
DO	-6	-23	3	**92**	
Hardness	**86**	27	-27	19	
TKN	36	32	**82**	5	
NO2 + NO3	**82**	-3	-16	7	
pH	21	27	-10	**85**	
Total P	**67**	29	51	-6	
TSS	-8	**90**	28	-1	
SC	**94**	-16	22	4	
Sulfate	**80**	7	29	-21	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-15	**93**	20	7	
Iron	-16	**92**	19	-7	Final Communality
Eigenvalue	4.873	3.263	3.223	1.727	13.087
% Variance Explained	32.4867	21.7533	21.4867	11.5133	87.24666667

Supplementary Table 2.18 – Quarter 4 Median Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**77**	-38	-37	16	
TOC	-3	3	**95**	-16	
Chloride	**86**	6	33	13	
COD	-2	40	**89**	0	
DO	-5	-40	0	**83**	
Hardness	**85**	-31	-27	15	
TKN	40	51	**69**	3	
NO2 + NO3	**78**	-5	-22	-2	
pH	17	21	-13	**89**	
Total P	**70**	42	39	0	
TSS	-4	**95**	15	3	
SC	**97**	-8	14	7	
Sulfate	**83**	23	24	-11	
Temperature	28	**73**	19	-1	
Turbidity	-16	**95**	12	1	
Iron	-25	**85**	18	-23	Final Communality
Eigenvalue	5.149	4.166	2.886	1.631	13.832
% Variance Explained	32.1813	26.0375	18.0375	10.1938	86.45

Supplementary Table 2.19 – Quarter 4 Trimmed Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**77**	-33	-38	20	
TOC	-8	8	**91**	-12	
Chloride	**83**	-16	39	8	
COD	-1	38	**90**	0	
DO	-7	-31	1	**89**	
Hardness	**86**	-28	-29	17	
TKN	40	34	**80**	3	
NO2 + NO3	**80**	-7	-18	7	
pH	20	25	-10	**87**	
Total P	**69**	29	47	-4	
TSS	0	**93**	23	-2	
SC	**96**	-15	18	5	
Sulfate	**82**	11	25	-20	
Temperature	DNL	DNL	DNL	DNL	
Turbidity	-16	**93**	22	4	
Iron	-22	**91**	16	-11	Final Communality
Eigenvalue	4.99	3.309	3.159	1.705	13.163
% Variance Explained	33.2667	22.06	21.06	11.3667	87.75333333

Supplementary Table 2.20 – Quarter 4 Geometric Mean Dataset Factor Loadings; "DNL" indicates that a variable did meet the 0.6 loading criterion on any factor

	Factor 1	Factor 2	Factor 3	Factor 4	
Alkalinity	**78**	-40	-33	13	
TOC	-5	5	**96**	-13	
Chloride	**86**	2	36	10	
COD	-1	37	**91**	0	
DO	-2	-29	3	**88**	
Hardness	**87**	-31	-25	14	
TKN	35	40	**79**	4	
NO2 + NO3	**83**	-3	-12	7	
pH	17	10	-10	**89**	
Total P	**67**	40	44	-1	
TSS	2	**96**	13	1	
SC	**96**	-5	17	6	
Sulfate	**83**	23	24	-14	
Temperature	24	**73**	31	-5	
Turbidity	-18	**93**	18	-1	
Iron	-21	**88**	14	-23	Final Communality
Variance Explained	5.136	3.973	3.138	1.707	13.955
	32.1	24.8313	19.6125	10.6688	87.21875

Hotelling's Pairwise Cluster Comparison Tests

Supplementary Table 3.1 – Pairwise Hotelling's p-values for the annual geometric mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	0.000	0.001	0.000	2.76E-05
Cluster 2		0	4.71E-09	0.000	1.92E-05
Cluster 3			0	2.80E-05	9.16E-07
Cluster 4				0	1.05E-05
Cluster 5					0

Supplementary Table 3.2 – Pairwise Hotelling's p-values for the annual mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 1	0	3.21E-08	8.38E-06	2.83E-06	2.01E-06	0.000
Cluster 2		0	0.000	5.09E-08	4.90E-05	0.001
Cluster 3			0	0.001	5.45E-06	0.001
Cluster 4				0	2.37E-06	0.001
Cluster 5					0	5.33E-05
Cluster 6						0

Supplementary Table 3.3 – Pairwise Hotelling's p-values for the annual median factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	0.001	0.000	0.001	6.50E-08
Cluster 2		0	1.78E-06	0.000	3.39E-05
Cluster 3			0	4.34E-06	2.22E-07
Cluster 4				0	0.001
Cluster 5					0

Supplementary Table 3.4 – Pairwise Hotelling's p-values for the annual trimmed mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	1.75E-05	9.45E-06	4.49E-05	6.93E-05
Cluster 2		0	0.001	0.002	0.001
Cluster 3			0	7.95E-09	7.20E-07
Cluster 4				0	5.96E-05
Cluster 5					0

Supplementary Table 3.5 – Pairwise Hotelling's p-values for the quarter 1 geometric mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	3.63E-07	0.001	7.89E-06	2.03E-09
Cluster 2		0	7.57E-05	0.000	1.06E-06
Cluster 3			0	0.0309	1.83E-07
Cluster 4				0	0.001
Cluster 5					0

Supplementary Table 3.6 – Pairwise Hotelling's p-values for the quarter 1 mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Cluster 1	0	0.015	Fail	0.003	Fail	0.002	0.030	Fail	Fail	0.007
Cluster 2		0	0.019	8.27E-07	0.039	0.000	0.027	0.094	0.027	0.013
Cluster 3			0	0.007	Fail	0.001	0.051	Fail	Fail	0.178
Cluster 4				0	9.64E-05	7.73E-06	3.51E-06	1.61E-05	0.000	3.90E-05
Cluster 5					0	0.002	0.0633	Fail	Fail	0.032
Cluster 6						0	0.002	0.107	0.011	0.002
Cluster 7							0	0.111	0.193	0.033
Cluster 8								0	Fail	0.0162
Cluster 9									0	0.0126
Cluster 10										0

Supplementary Table 3.7 – Pairwise Hotelling's p-values for the quarter 1 median factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Cluster 1	0	0.000	0.120	0.004	0.053	Fail	Fail	Fail	0.067
Cluster 2		0	0.002	3.67E-08	0.000	0.000	0.012	0.017	0.000
Cluster 3			0	4.96E-05	0.169	Fail	Fail	Fail	0.033
Cluster 4				0	0.000	0.0016	2.53E-05	3.61E-05	0.001
Cluster 5					0	0.117	0.448	0.090	0.002
Cluster 6						0	Fail	Fail	0.141
Cluster 7							0	Fail	0.118
Cluster 8								0	0.164
Cluster 9									0

Supplementary Table 3.8 – Pairwise Hotelling's p-values for the quarter 1 trimmed mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster 1	0	0.000	0.001	0.004	0.000	1.90E-05	0.000	1.35E-05
Cluster 2		0	Fail	Fail	0.013	0.090	0.040	1.77E-05
Cluster 3			0	Fail	0.003	0.015	0.023	7.86E-05
Cluster 4				0	0.001	0.013	0.018	0.006
Cluster 5					0	0.003	0.001	9.23E-07
Cluster 6						0	0.013	0.000
Cluster 7							0	5.51E-08
Cluster 8								0

Supplementary Table 3.9 – Pairwise Hotelling's p-values for the quarter 2 geometric mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	0	0.000	0.000	0.061	0.001	0.090	0.000
Cluster 2		0	2.82E-06	2.65E-05	0.000	8.43E-06	4.60E-05
Cluster 3			0	0.001	0.001	0.002	3.21E-05
Cluster 4				0	0.001	Fail	0.001
Cluster 5					0	0.002	1.26E-05
Cluster 6						0	3.33E-05
Cluster 7							0

Supplementary Table 3.10 – Pairwise Hotelling's p-values for the quarter 2 mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	1.19E-05	2.28E-06	0.001	8.06E-06
Cluster 2		0	1.92E-06	2.15E-06	0.000
Cluster 3			0	6.07E-06	5.90E-06
Cluster 4				0	0.001
Cluster 5					0

Supplementary Table 3.11 – Pairwise Hotelling's p-values for the quarter 2 median factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	2.40E-05	3.01E-06	0.003	7.09E-05
Cluster 2		0	3.05E-05	4.75E-08	1.75E-05
Cluster 3			0	9.63E-06	0.000
Cluster 4				0	1.95E-06
Cluster 5					0

Supplementary Table 3.12 – Pairwise Hotelling's p-values for the quarter 2 trimmed mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	0	3.36E-05	0.001	3.16E-06	0.000	8.60E-05	1.38E-05
Cluster 2		0	4.75E-06	0.147	0.065	0.000	0.005
Cluster 3			0	0.000	0.001	3.01E-06	4.75E-05
Cluster 4				0	0.033	4.11E-05	0.000
Cluster 5					0	0.000	0.000
Cluster 6						0	4.53E-05
Cluster 7							0

Supplementary Table 3.13 – Pairwise Hotelling's p-values for the quarter 3 geometric mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	5.67E-06	7.74E-06	2.42E-06	0.000
Cluster 2		0	6.75E-05	7.14E-05	3.66E-05
Cluster 3			0	0.000	0.000
Cluster 4				0	0.001
Cluster 5					0

Supplementary Table 3.14 – Pairwise Hotelling's p-values for the quarter 3 mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	0	3.94E-05	0.003	0.037	0.000	4.84E-05	0.003
Cluster 2		0	1.52E-06	0.001	0.000	0.000	1.04E-06
Cluster 3			0	0.022	0.002	0.000	0.004
Cluster 4				0	0.003	0.003	0.006
Cluster 5					0	0.006	0.021
Cluster 6						0	0.004
Cluster 7							0

Supplementary Table 3.15 – Pairwise Hotelling's p-values for the quarter 3 median factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	4.45E-06	1.29E-06	9.96E-06	3.89E-06
Cluster 2		0	8.12E-08	1.58E-05	0.000
Cluster 3			0	2.57E-05	0.000
Cluster 4				0	0.000
Cluster 5					0

Supplementary Table 3.16 – Pairwise Hotelling's p-values for the quarter 3 trimmed mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	0.001	1.13E-06	0.000	2.12E-06
Cluster 2		0	2.02E-05	0.002	0.000
Cluster 3			0	0.000	1.20E-06
Cluster 4				0	2.54E-06
Cluster 5					0

Supplementary Table 3.17 – Pairwise Hotelling's p-values for the quarter 4 geometric mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	0	1.89E-05	4.19E-05	0.004	2.90E-05	0.017	2.57E-06
Cluster 2		0	8.45E-08	0.005	1.22E-05	0.001	0.000
Cluster 3			0	3.18E-06	8.71E-05	0.000	0.001
Cluster 4				0	1.97E-06	0.120	0.001
Cluster 5					0	0.001	7.93E-06
Cluster 6						0	0.007
Cluster 7							0

Supplementary Table 3.18 – Pairwise Hotelling's p-values for the quarter 4 mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	3.81E-03	4.39E-04	3.32E-06	0.001
Cluster 2		0	0.000	6.30E-07	2.38E-04
Cluster 3			0	0.000	2.29E-07
Cluster 4				0	0.000
Cluster 5					0

Supplementary Table 3.19 – Pairwise Hotelling's p-values for the quarter 4 median factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Cluster 1	0	3.78E-06	0.001	0.000	0.001	0.133	0.000
Cluster 2		0	2.99E-05	2.28E-05	1.67E-05	0.002	0.001
Cluster 3			0	0.002	0.001	0.066	0.000
Cluster 4				0	0.000	0.027	0.001
Cluster 5					0	0.001	3.11E-06
Cluster 6						0	0.047
Cluster 7							0

Supplementary Table 3.20 – Pairwise Hotelling's p-values for the quarter 4 trimmed mean factor clusters; tests that indicate clusters are not different at an α=0.05 significance level and tests that fail due to lack of samples are highlighted

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 1	0	5.46E-05	5.22E-07	8.34E-09	5.15E-06	1.43E-06
Cluster 2		0	4.57E-07	3.20E-05	0.002	0.001
Cluster 3			0	0.001	1.18E-05	4.11E-07
Cluster 4				0	0.008	0.001
Cluster 5					0	0.000
Cluster 6						0

IDEM Station Cluster Assignments

Supplementary Table 4.1 – Annual Factor Cluster Assignments

Station name	Geomean (5)	Mean (6)	Median (5)	Trimmed Mean (5)
BL-.7	5	5	3	1
BL-64	2	2	1	3
BWC-4	5	5	3	5
CIC-17	2	5	1	1
EC-1	4	3	2	2
EC-21	2	2	1	4
EC-7	4	3	2	4
EEL-1	3	1	5	3
EEL-38	3	1	5	3
EW-1	3	1	5	1
EW-168	5	1	3	1
EW-239	5	5	3	1
EW-79	3	1	5	3
EW-94	3	1	5	1
FC-0.6	4	3	2	2
FC-26	5	5	3	4
FC-7	4	3	2	2
FR-17	5	5	3	1
FR-64	5	5	3	4
GC-8	4	3	2	5
IN-2	3	4	5	3
IWC-9	2	2	1	2
LST-2	3	4	5	5
MC-18	3	1	5	3
MC-35	5	5	3	4
MU-20	3	4	5	4
SGR-1	5	5	3	5
SLT-12	3	4	5	3
SND-4	3	1	5	5
VF-38	4	3	2	3
WLC-2	5	5	3	4
WR-134	1	6	4	5
WR-162	1	6	4	4
WR-19	1	6	4	3
WR-192	2	2	1	4
WR-210	2	2	1	4
WR-248	2	2	1	2
WR-279	2	6	3	5
WR-293	2	2	1	1
WR-309	2	2	1	1
WR-319	2	1	1	3
WR-348	5	5	3	3
WR-46	1	6	4	1
WR-81	1	6	4	3

Supplementary Table 4.2 – Quarter 1 Factor Cluster Assignments

Station name	Geomean (5)	Mean (10)	Median (9)	Trimmed Mean (8)
BL-.7	2	7	9	1
BL-64	5	2	3	7
BWC-4	2	10	5	1
CIC-17	2	10	1	1
EC-1	5	8	4	6
EC-21	5	2	4	7
EC-7	4	8	4	2
EEL-1	1	4	2	8
EEL-38	1	4	2	8
EW-1	1	4	2	8
EW-168	1	10	2	8
EW-239	2	10	9	1
EW-79	1	4	2	8
EW-94	1	4	2	8
FC-0.6	5	6	4	5
FC-26	5	2	5	7
FC-7	5	2	3	7
FR-17	2	7	5	1
FR-64	2	2	5	7
GC-8	4	9	6	2
IN-2	3	3	7	4
IWC-9	5	6	4	5
LST-2	3	3	7	4
MC-18	2	10	1	1
MC-35	2	2	1	7
MU-20	3	5	8	3
SGR-1	2	7	9	1
SLT-12	3	5	8	3
SND-4	2	4	5	8
VF-38	4	9	6	2
WLC-2	2	6	9	5
WR-134	1	4	2	8
WR-162	5	1	2	6
WR-19	1	4	2	8
WR-192	5	1	3	6
WR-210	5	1	3	6
WR-248	5	6	4	5
WR-279	5	6	4	5
WR-293	5	6	4	5
WR-309	5	6	4	5
WR-319	5	2	1	7
WR-348	2	7	9	1
WR-46	1	4	2	8
WR-81	1	4	2	8

Supplementary Table 4.3 – Quarter 2 Factor Cluster Assignments

Station name	Geomean (7)	Mean (5)	Median (5)	Trimmed Mean (7)
BL-.7	7	3	2	6
BL-64	2	1	3	1
BWC-4	7	3	2	6
CIC-17	2	3	2	1
EC-1	1	4	5	5
EC-21	2	1	3	1
EC-7	1	4	5	5
EEL-1	3	2	4	7
EEL-38	3	2	4	7
EW-1	3	2	4	7
EW-168	3	3	2	7
EW-239	7	3	2	6
EW-79	3	5	1	7
EW-94	3	2	4	7
FC-0.6	1	4	5	5
FC-26	2	3	2	1
FC-7	1	4	5	5
FR-17	7	3	2	6
FR-64	7	3	2	6
GC-8	4	4	5	2
IN-2	6	5	1	4
IWC-9	2	1	3	1
LST-2	3	5	1	4
MC-18	3	3	2	7
MC-35	7	3	5	6
MU-20	6	5	1	4
SGR-1	7	3	2	6
SLT-12	6	5	1	4
SND-4	3	3	2	7
VF-38	4	4	5	2
WLC-2	7	4	2	6
WR-134	5	2	4	3
WR-162	5	2	4	3
WR-19	5	2	4	3
WR-192	5	1	3	3
WR-210	2	1	3	1
WR-248	2	1	3	1
WR-279	2	1	3	1
WR-293	2	1	3	1
WR-309	2	1	3	1
WR-319	2	1	3	1
WR-348	7	3	2	6
WR-46	5	2	4	3
WR-81	5	2	4	3

Supplementary Table 4.4 – Quarter 3 Factor Cluster Assignments

Station name	Geomean (5)	Mean (7)	Median (5)	Trimmed Mean (5)
BL-.7	4	1	2	3
BL-64	2	2	1	5
BWC-4	4	1	2	3
CIC-17	4	2	1	5
EC-1	5	5	4	2
EC-21	2	2	1	5
EC-7	5	5	4	2
EEL-1	1	7	3	1
EEL-38	1	7	2	1
EW-1	1	7	3	1
EW-168	4	1	2	3
EW-239	4	1	2	3
EW-79	1	7	3	1
EW-94	1	7	3	1
FC-0.6	5	5	3	2
FC-26	4	1	2	3
FC-7	5	5	3	2
FR-17	4	4	2	3
FR-64	4	4	2	3
GC-8	5	5	4	2
IN-2	3	3	5	4
IWC-9	2	2	1	5
LST-2	3	3	5	4
MC-18	4	1	2	3
MC-35	5	4	4	3
MU-20	3	3	5	4
SGR-1	4	4	2	3
SLT-12	3	3	5	4
SND-4	4	3	5	4
VF-38	5	5	4	2
WLC-2	4	1	2	3
WR-134	1	6	3	1
WR-162	1	6	3	1
WR-19	1	6	3	1
WR-192	2	2	1	5
WR-210	2	2	1	5
WR-248	2	2	1	5
WR-279	2	2	1	5
WR-293	2	2	1	5
WR-309	2	2	1	5
WR-319	2	1	2	3
WR-348	4	1	4	3
WR-46	1	6	3	1
WR-81	1	6	3	1

Supplementary Table 4.5 – Quarter 4 Factor Cluster Assignments

Station name	Geomean (7)	Mean (5)	Median (7)	Trimmed Mean (6)
BL-.7	5	1	2	1
BL-64	3	1	7	2
BWC-4	5	2	2	3
CIC-17	1	1	7	1
EC-1	1	4	4	5
EC-21	3	5	7	2
EC-7	1	4	4	5
EEL-1	2	2	5	3
EEL-38	2	2	5	3
EW-1	2	2	5	3
EW-168	2	1	5	3
EW-239	5	1	2	1
EW-79	2	3	5	3
EW-94	2	3	5	3
FC-0.6	1	4	4	5
FC-26	5	1	2	1
FC-7	1	4	4	5
FR-17	5	1	2	1
FR-64	5	1	2	3
GC-8	6	4	4	5
IN-2	4	3	1	4
IWC-9	3	5	7	2
LST-2	4	3	1	4
MC-18	2	3	5	3
MC-35	5	1	2	1
MU-20	4	3	1	4
SGR-1	5	1	2	1
SLT-12	4	3	1	4
SND-4	1	2	4	3
VF-38	6	4	4	5
WLC-2	5	1	2	1
WR-134	7	2	3	6
WR-162	7	5	3	6
WR-19	7	2	3	6
WR-192	3	5	6	2
WR-210	3	5	6	2
WR-248	3	5	7	2
WR-279	3	1	7	1
WR-293	3	1	7	1
WR-309	3	5	7	2
WR-319	1	1	7	3
WR-348	1	1	4	1
WR-46	7	2	3	6
WR-81	7	2	3	6

Supplementary Table 4.6 – Annual SOM Cluster Assignments

Station name	Geomean (8)	Mean (3)	Median (8)	Trimmed Mean (7)
BL-.7	5	1	2	4
BL-64	4	1	4	6
BWC-4	5	2	2	4
CIC-17	4	1	4	6
EC-1	3	1	6	3
EC-21	4	1	4	6
EC-7	3	1	6	3
EEL-1	1	2	7	2
EEL-38	1	2	7	2
EW-1	1	2	7	2
EW-168	5	2	2	4
EW-239	5	1	2	4
EW-79	1	2	7	2
EW-94	1	2	7	2
FC-0.6	3	1	6	3
FC-26	5	1	2	4
FC-7	3	1	6	3
FR-17	5	1	2	4
FR-64	5	1	2	4
GC-8	2	2	5	7
IN-2	8	2	3	7
IWC-9	4	1	4	6
LST-2	8	2	3	7
MC-18	5	2	2	4
MC-35	5	1	2	4
MU-20	8	2	3	7
SGR-1	5	1	2	4
SLT-12	8	2	3	7
SND-4	2	2	5	7
VF-38	2	2	5	7
WLC-2	4	1	4	6
WR-134	6	3	1	5
WR-162	6	3	8	1
WR-19	6	3	1	5
WR-192	7	3	8	1
WR-210	7	3	8	1
WR-248	4	1	4	6
WR-279	4	3	4	6
WR-293	4	1	4	6
WR-309	4	1	4	6
WR-319	5	1	2	4
WR-348	5	1	2	4
WR-46	6	3	1	5
WR-81	6	3	1	5

Supplementary Table 4.7 – Quarter 1 SOM Cluster Assignments

Station Name	Geomean (9)	Mean (7)	Median (6)	Trimmed Mean (6)
BL-.7	1	7	3	5
BL-64	7	7	2	1
BWC-4	1	6	3	5
CIC-17	1	6	3	5
EC-1	6	5	4	1
EC-21	7	7	2	5
EC-7	6	5	4	1
EEL-1	5	2	5	4
EEL-38	5	2	5	4
EW-1	5	2	5	4
EW-168	1	6	3	5
EW-239	1	7	3	5
EW-79	5	3	6	4
EW-94	5	2	6	4
FC-0.6	6	5	4	1
FC-26	1	7	3	5
FC-7	4	7	4	5
FR-17	1	7	3	5
FR-64	1	7	3	5
GC-8	2	3	6	6
IN-2	3	3	6	6
IWC-9	7	5	2	1
LST-2	3	3	6	6
MC-18	1	6	3	5
MC-35	1	7	3	5
MU-20	5	3	6	4
SGR-1	1	7	3	5
SLT-12	3	3	6	6
SND-4	2	6	3	4
VF-38	2	3	6	6
WLC-2	7	5	2	1
WR-134	8	1	1	3
WR-162	9	4	1	2
WR-19	8	1	5	3
WR-192	9	4	1	2
WR-210	9	4	1	2
WR-248	7	5	1	1
WR-279	7	5	2	1
WR-293	7	5	2	1
WR-309	7	5	2	1
WR-319	4	7	3	5
WR-348	1	7	3	5
WR-46	8	1	5	3
WR-81	8	1	5	3

Supplementary Table 4.8 – Quarter 2 SOM Cluster Assignments

Station Name	Geomean (9)	Mean (9)	Median (8)	Trimmed Mean (6)
BL-.7	3	8	4	2
BL-64	3	6	4	2
BWC-4	2	5	2	2
CIC-17	3	6	4	2
EC-1	6	2	6	4
EC-21	3	6	4	2
EC-7	6	2	6	4
EEL-1	5	7	5	6
EEL-38	5	7	5	3
EW-1	5	7	5	3
EW-168	2	5	2	2
EW-239	3	8	2	2
EW-79	5	4	5	3
EW-94	5	7	5	3
FC-0.6	6	2	6	4
FC-26	3	6	4	2
FC-7	2	2	6	4
FR-17	3	8	2	2
FR-64	3	8	2	2
GC-8	4	4	3	3
IN-2	1	1	1	1
IWC-9	7	3	7	5
LST-2	1	1	1	1
MC-18	2	5	2	2
MC-35	3	8	2	2
MU-20	4	4	5	3
SGR-1	3	5	4	2
SLT-12	1	1	1	1
SND-4	2	5	2	3
VF-38	4	4	3	3
WLC-2	6	2	4	4
WR-134	8	9	8	6
WR-162	8	9	8	6
WR-19	8	9	8	6
WR-192	9	3	7	5
WR-210	9	3	7	5
WR-248	7	3	7	5
WR-279	7	3	7	5
WR-293	7	3	7	5
WR-309	7	3	7	5
WR-319	2	5	2	2
WR-348	3	8	2	2
WR-46	8	9	8	6
WR-81	8	9	8	6

Supplementary Table 4.9 – Quarter 3 SOM Cluster Assignments

Station Name	Geomean (7)	Mean (4)	Median (5)	Trimmed Mean (4)
BL-.7	7	1	1	4
BL-64	3	1	3	4
BWC-4	7	1	1	4
CIC-17	3	3	3	4
EC-1	5	3	4	1
EC-21	3	3	3	4
EC-7	5	2	2	3
EEL-1	4	2	2	3
EEL-38	6	2	1	3
EW-1	4	2	5	3
EW-168	7	1	1	4
EW-239	7	1	1	4
EW-79	6	2	2	3
EW-94	4	2	2	3
FC-0.6	5	2	2	1
FC-26	7	1	1	4
FC-7	5	2	2	3
FR-17	7	1	1	4
FR-64	7	1	1	4
GC-8	6	2	2	3
IN-2	6	2	2	3
IWC-9	3	3	3	1
LST-2	6	2	2	3
MC-18	7	1	1	4
MC-35	7	1	1	4
MU-20	6	2	2	3
SGR-1	7	1	1	4
SLT-12	6	2	2	3
SND-4	6	2	2	3
VF-38	6	2	2	3
WLC-2	7	1	1	4
WR-134	1	4	5	2
WR-162	2	4	4	1
WR-19	1	4	5	2
WR-192	2	3	4	1
WR-210	2	3	4	1
WR-248	2	3	3	1
WR-279	3	3	3	1
WR-293	3	3	3	4
WR-309	3	3	3	1
WR-319	7	1	1	4
WR-348	7	1	1	4
WR-46	1	4	5	2
WR-81	1	4	5	2

Supplementary Table 4.10 – Quarter 4 SOM Cluster Assignments

Station Name	Geomean (5)	Mean (5)	Median (6)	Trimmed Mean (7)
BL-.7	5	3	6	2
BL-64	1	3	6	2
BWC-4	5	3	6	2
CIC-17	1	2	6	1
EC-1	2	2	5	1
EC-21	1	2	5	1
EC-7	4	4	4	4
EEL-1	4	5	2	6
EEL-38	4	5	2	6
EW-1	4	5	2	6
EW-168	5	3	6	2
EW-239	5	3	6	2
EW-79	4	4	2	6
EW-94	4	4	2	6
FC-0.6	4	4	4	4
FC-26	5	3	6	2
FC-7	4	4	4	4
FR-17	5	3	6	2
FR-64	5	3	6	2
GC-8	4	4	4	4
IN-2	4	4	3	7
IWC-9	1	2	5	1
LST-2	4	4	3	7
MC-18	5	3	6	4
MC-35	5	3	6	2
MU-20	4	4	3	7
SGR-1	5	3	6	2
SLT-12	4	4	3	7
SND-4	4	4	4	4
VF-38	4	4	4	4
WLC-2	1	3	6	2
WR-134	2	1	1	3
WR-162	2	1	1	3
WR-19	3	5	1	5
WR-192	2	1	1	3
WR-210	2	1	1	3
WR-248	1	2	5	1
WR-279	1	2	5	1
WR-293	1	2	5	1
WR-309	1	1	5	1
WR-319	5	3	6	2
WR-348	5	3	6	2
WR-46	3	5	1	5
WR-81	3	5	1	5

Cluster Comparison T-tests

Supplementary Table 5.1 – The Annual Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Mean Factors	M/V	HIGH	LOW	LOW	HIGH	M/V				
Total Organic Carbon	Annual Mean Factors	M/V	M/V	HIGH	M/V	LOW	HIGH				
Chloride	Annual Mean Factors	LOW	HIGH	M/V	LOW	M/V	M/V				
Chemical Oxygen Demand	Annual Mean Factors	M/V	M/V	HIGH	M/V	LOW	HIGH				
Dissolved Oxygen	Annual Mean Factors	LOW	M/V	HIGH	LOW	M/V	M/V				
Hardness	Annual Mean Factors	M/V	HIGH	LOW	LOW	HIGH	M/V				
Total Kjeldahl Nitrogen	Annual Mean Factors	M/V	HIGH	M/V	LOW	LOW	HIGH				
Nitrate + Nitrite	Annual Mean Factors	M/V	HIGH	LOW	LOW	HIGH	M/V				
pH	Annual Mean Factors	M/V	M/V	M/V	LOW	HIGH	HIGH				
Total Phosphorus	Annual Mean Factors	M/V	HIGH	M/V	LOW	M/V	HIGH				
Total Suspended Solids	Annual Mean Factors	HIGH	M/V	LOW	M/V	M/V	HIGH				
Specific Conductance	Annual Mean Factors	LOW	HIGH	M/V	LOW	M/V	M/V				
Sulfate	Annual Mean Factors	LOW	HIGH	M/V	M/V	LOW	HIGH				
Temperature	Annual Mean Factors	M/V	M/V	M/V	M/V	LOW	HIGH				
Turbidity	Annual Mean Factors	HIGH	LOW	LOW	M/V	M/V	HIGH				
Iron	Annual Mean Factors	HIGH	M/V	LOW	M/V	M/V	HIGH				

Supplementary Table 5.2 – The Annual Median Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Median Factors	HIGH	M/V	HIGH	M/V	LOW					
Total Organic Carbon	Annual Median Factors	M/V	HIGH	LOW	HIGH	M/V					
Chloride	Annual Median Factors	HIGH	M/V	M/V	M/V	LOW					
Chemical Oxygen Demand	Annual Median Factors	M/V	HIGH	LOW	HIGH	M/V					
Dissolved Oxygen	Annual Median Factors	M/V	HIGH	M/V	M/V	LOW					
Hardness	Annual Median Factors	HIGH	LOW	HIGH	M/V	LOW					
Total Kjeldahl Nitrogen	Annual Median Factors	M/V	M/V	LOW	HIGH	M/V					
Nitrate + Nitrite	Annual Median Factors	HIGH	LOW	M/V	M/V	LOW					
pH	Annual Median Factors	M/V	M/V	HIGH	HIGH	LOW					
Total Phosphorus	Annual Median Factors	HIGH	M/V	M/V	HIGH	M/V					
Total Suspended Solids	Annual Median Factors	M/V	M/V	M/V	HIGH	M/V					
Specific Conductance	Annual Median Factors	HIGH	M/V	M/V	M/V	LOW					
Sulfate	Annual Median Factors	HIGH	M/V	M/V	HIGH	LOW					
Temperature	Annual Median Factors	M/V	M/V	M/V	HIGH	M/V					
Turbidity	Annual Median Factors	M/V	M/V	M/V	HIGH	HIGH					
Iron	Annual Median Factors	M/V	LOW	LOW	HIGH	HIGH					

Supplementary Table 5.3 – The Annual Trimmed Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Trimmed Mean Factors	HIGH	M/V	LOW	HIGH	M/V					
Total Organic Carbon	Annual Trimmed Mean Factors	LOW	HIGH	M/V	M/V	M/V					
Chloride	Annual Trimmed Mean Factors	M/V	M/V	LOW	HIGH	M/V					
Chemical Oxygen Demand	Annual Trimmed Mean Factors	LOW	HIGH	M/V	M/V	M/V					
Dissolved Oxygen	Annual Trimmed Mean Factors	M/V	M/V	LOW	M/V	HIGH					
Hardness	Annual Trimmed Mean Factors	HIGH	M/V	LOW	HIGH	M/V					
Total Kjeldahl Nitrogen	Annual Trimmed Mean Factors	LOW	HIGH	M/V	HIGH	M/V					
Nitrate + Nitrite	Annual Trimmed Mean Factors	HIGH	M/V	LOW	HIGH	LOW					
pH	Annual Trimmed Mean Factors	HIGH	HIGH	LOW	M/V	M/V					
Total Phosphorus	Annual Trimmed Mean Factors	M/V	HIGH	M/V	HIGH	LOW					
Total Suspended Solids	Annual Trimmed Mean Factors	M/V	HIGH	HIGH	M/V	LOW					
Specific Conductance	Annual Trimmed Mean Factors	M/V	M/V	LOW	HIGH	M/V					
Sulfate	Annual Trimmed Mean Factors	LOW	HIGH	LOW	HIGH	M/V					
Temperature	Annual Trimmed Mean Factors	M/V	HIGH	M/V	M/V	M/V					
Turbidity	Annual Trimmed Mean Factors	M/V	HIGH	HIGH	M/V	LOW					
Iron	Annual Trimmed Mean Factors	M/V	HIGH	HIGH	M/V	LOW					

Supplementary Table 5.4 – The Annual Geometric Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Geomean Factors	M/V	HIGH	LOW	M/V	HIGH					
Total Organic Carbon	Annual Geomean Factors	HIGH	M/V	M/V	HIGH	LOW					
Chloride	Annual Geomean Factors	M/V	HIGH	LOW	M/V	M/V					
Chemical Oxygen Demand	Annual Geomean Factors	HIGH	M/V	M/V	HIGH	LOW					
Dissolved Oxygen	Annual Geomean Factors	M/V	M/V	LOW	HIGH	M/V					
Hardness	Annual Geomean Factors	M/V	HIGH	LOW	LOW	HIGH					
Total Kjeldahl Nitrogen	Annual Geomean Factors	HIGH	M/V	M/V	M/V	LOW					
Nitrate + Nitrite	Annual Geomean Factors	M/V	HIGH	LOW	LOW	M/V					
pH	Annual Geomean Factors	HIGH	M/V	LOW	M/V	HIGH					
Total Phosphorus	Annual Geomean Factors	HIGH	HIGH	M/V	M/V	M/V					
Total Suspended Solids	Annual Geomean Factors	HIGH	M/V	M/V	M/V	M/V					
Specific Conductance	Annual Geomean Factors	M/V	HIGH	LOW	M/V	M/V					
Sulfate	Annual Geomean Factors	HIGH	HIGH	LOW	M/V	LOW					
Temperature	Annual Geomean Factors	HIGH	M/V	M/V	M/V	M/V					
Turbidity	Annual Geomean Factors	HIGH	M/V	HIGH	M/V	LOW					
Iron	Annual Geomean Factors	HIGH	M/V	HIGH	LOW	LOW					

Supplementary Table 5.5 – The Quarter 1 Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Mean Factors	M/V	HIGH	LOW	LOW	LOW	HIGH	M/V	M/V	LOW	M/V
Total Organic Carbon	Q1 Mean Factors	HIGH	LOW	LOW	HIGH	M/V	M/V	LOW	HIGH	HIGH	M/V
Chloride	Q1 Mean Factors	HIGH	M/V	LOW	M/V	LOW	HIGH	M/V	HIGH	LOW	M/V
Chemical Oxygen Demand	Q1 Mean Factors	HIGH	LOW	LOW	HIGH	M/V	HIGH	LOW	M/V	M/V	M/V
Dissolved Oxygen	Q1 Mean Factors	LOW	M/V	M/V	M/V	M/V	M/V	HIGH	M/V	HIGH	M/V
Hardness	Q1 Mean Factors	M/V	HIGH	LOW	LOW	LOW	HIGH	M/V	M/V	LOW	M/V
Total Kjeldahl Nitrogen	Q1 Mean Factors	HIGH	LOW	LOW	M/V	M/V	HIGH	LOW	M/V	M/V	M/V
Nitrate + Nitrite	Q1 Mean Factors	M/V	M/V	LOW	M/V	LOW	M/V	HIGH	M/V	LOW	HIGH
pH	Q1 Mean Factors	LOW	M/V	LOW	M/V	LOW	HIGH	HIGH	LOW	M/V	M/V
Total Phosphorus	Q1 Mean Factors	HIGH	LOW	LOW	HIGH	M/V	HIGH	M/V	M/V	M/V	M/V
Total Suspended Solids	Q1 Mean Factors	HIGH	LOW	M/V	HIGH	M/V	M/V	LOW	M/V	M/V	M/V
Specific Conductance	Q1 Mean Factors	HIGH	M/V	LOW	M/V	LOW	HIGH	M/V	M/V	LOW	M/V
Sulfate	Q1 Mean Factors	HIGH	M/V	M/V	M/V	LOW	HIGH	M/V	M/V	LOW	LOW
Temperature	Q1 Mean Factors	HIGH	M/V	M/V	HIGH	M/V	LOW	M/V	M/V	M/V	M/V
Turbidity	Q1 Mean Factors	M/V	LOW	M/V	HIGH	M/V	M/V	M/V	M/V	M/V	M/V
Iron	Q1 Mean Factors	M/V	LOW	M/V	HIGH	M/V	M/V	M/V	LOW	M/V	M/V

Supplementary Table 5.6 – The Quarter 1 Median Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Median Factors	M/V	LOW	M/V	HIGH	M/V	LOW	LOW	LOW	HIGH	
Total Organic Carbon	Q1 Median Factors	M/V	M/V	M/V	HIGH	LOW	M/V	LOW	M/V	LOW	
Chloride	Q1 Median Factors	M/V	M/V	HIGH	HIGH	M/V	LOW	LOW	LOW	M/V	
Chemical Oxygen Demand	Q1 Median Factors	LOW	HIGH	M/V	M/V	LOW	M/V	LOW	M/V	LOW	
Dissolved Oxygen	Q1 Median Factors	M/V	LOW	LOW	M/V	M/V	M/V	M/V	M/V	HIGH	
Hardness	Q1 Median Factors	M/V	LOW	M/V	M/V	M/V	LOW	LOW	LOW	HIGH	
Total Kjeldahl Nitrogen	Q1 Median Factors	LOW	HIGH	HIGH	M/V	LOW	M/V	LOW	M/V	LOW	
Nitrate + Nitrite	Q1 Median Factors	M/V	M/V	M/V	M/V	HIGH	LOW	M/V	LOW	M/V	
pH	Q1 Median Factors	M/V	LOW	M/V	HIGH	HIGH	M/V	LOW	LOW	HIGH	
Total Phosphorus	Q1 Median Factors	M/V	HIGH	M/V	M/V	LOW	M/V	LOW	M/V	M/V	
Total Suspended Solids	Q1 Median Factors	LOW	HIGH	M/V	LOW	M/V	LOW	M/V	M/V	M/V	
Specific Conductance	Q1 Median Factors	M/V	M/V	HIGH	HIGH	M/V	LOW	LOW	LOW	M/V	
Sulfate	Q1 Median Factors	M/V	M/V	HIGH	HIGH	M/V	LOW	M/V	LOW	M/V	
Temperature	Q1 Median Factors	LOW	M/V	HIGH	LOW	HIGH	M/V	HIGH	M/V	LOW	
Turbidity	Q1 Median Factors	LOW	HIGH	M/V	M/V	LOW	M/V	M/V	M/V	LOW	
Iron	Q1 Median Factors	M/V	HIGH	M/V	LOW	M/V	M/V	M/V	M/V	LOW	

Supplementary Table 5.7 – The Quarter 1 Trimmed Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Trimmed Mean Factors	HIGH	LOW	LOW	LOW	HIGH	M/V	HIGH	LOW		
Total Organic Carbon	Q1 Trimmed Mean Factors	LOW	HIGH	M/V	LOW	M/V	HIGH	M/V	M/V		
Chloride	Q1 Trimmed Mean Factors	M/V	M/V	LOW	LOW	HIGH	HIGH	M/V	M/V		
Chemical Oxygen Demand	Q1 Trimmed Mean Factors	LOW	HIGH	M/V	LOW	M/V	HIGH	LOW	HIGH		
Dissolved Oxygen	Q1 Trimmed Mean Factors	HIGH	HIGH	M/V	M/V	M/V	M/V	M/V	M/V		
Hardness	Q1 Trimmed Mean Factors	HIGH	LOW	LOW	LOW	HIGH	M/V	HIGH	LOW		
Total Kjeldahl Nitrogen	Q1 Trimmed Mean Factors	LOW	M/V	M/V	LOW	HIGH	HIGH	LOW	HIGH		
Nitrate + Nitrite	Q1 Trimmed Mean Factors	HIGH	LOW	LOW	LOW	M/V	M/V	M/V	M/V		
pH	Q1 Trimmed Mean Factors	HIGH	M/V	LOW	LOW	HIGH	M/V	M/V	M/V		
Total Phosphorus	Q1 Trimmed Mean Factors	M/V	M/V	M/V	LOW	HIGH	HIGH	LOW	HIGH		
Total Suspended Solids	Q1 Trimmed Mean Factors	M/V	M/V	M/V	M/V	M/V	M/V	LOW	HIGH		
Specific Conductance	Q1 Trimmed Mean Factors	M/V	LOW	LOW	LOW	HIGH	HIGH	M/V	M/V		
Sulfate	Q1 Trimmed Mean Factors	LOW	LOW	LOW	M/V	HIGH	HIGH	M/V	M/V		
Temperature	Q1 Trimmed Mean Factors	LOW	M/V	M/V	M/V	LOW	HIGH	M/V	M/V		
Turbidity	Q1 Trimmed Mean Factors	M/V	M/V	M/V	M/V	M/V	M/V	LOW	HIGH		
Iron	Q1 Trimmed Mean Factors	LOW	M/V	M/V	M/V	M/V	M/V	LOW	HIGH		

Supplementary Table 5.8 – The Quarter 1 Geometric Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Geomean Factors	LOW	HIGH	LOW	LOW	HIGH					
Total Organic Carbon	Q1 Geomean Factors	M/V	LOW	M/V	HIGH	HIGH					
Chloride	Q1 Geomean Factors	M/V	M/V	LOW	M/V	HIGH					
Chemical Oxygen Demand	Q1 Geomean Factors	HIGH	LOW	M/V	M/V	M/V					
Dissolved Oxygen	Q1 Geomean Factors	LOW	HIGH	M/V	HIGH	LOW					
Hardness	Q1 Geomean Factors	LOW	HIGH	LOW	LOW	HIGH					
Total Kjeldahl Nitrogen	Q1 Geomean Factors	M/V	LOW	LOW	M/V	HIGH					
Nitrate + Nitrite	Q1 Geomean Factors	M/V	HIGH	LOW	LOW	M/V					
pH	Q1 Geomean Factors	LOW	HIGH	LOW	M/V	M/V					
Total Phosphorus	Q1 Geomean Factors	HIGH	M/V	LOW	M/V	M/V					
Total Suspended Solids	Q1 Geomean Factors	HIGH	LOW	M/V	LOW	M/V					
Specific Conductance	Q1 Geomean Factors	M/V	M/V	LOW	LOW	HIGH					
Sulfate	Q1 Geomean Factors	M/V	LOW	LOW	LOW	HIGH					
Temperature	Q1 Geomean Factors	M/V	LOW	M/V	M/V	M/V					
Turbidity	Q1 Geomean Factors	HIGH	LOW	M/V	M/V	M/V					
Iron	Q1 Geomean Factors	HIGH	LOW	M/V	M/V	LOW					

Supplementary Table 5.9 – The Quarter 2 Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Mean Factors	HIGH	LOW	HIGH	M/V	LOW					
Total Organic Carbon	Q2 Mean Factors	M/V	M/V	LOW	M/V	M/V					
Chloride	Q2 Mean Factors	HIGH	M/V	M/V	M/V	LOW					
Chemical Oxygen Demand	Q2 Mean Factors	M/V	HIGH	LOW	M/V	M/V					
Dissolved Oxygen	Q2 Mean Factors	M/V	M/V	M/V	HIGH	LOW					
Hardness	Q2 Mean Factors	HIGH	LOW	HIGH	M/V	LOW					
Total Kjeldahl Nitrogen	Q2 Mean Factors	M/V	HIGH	LOW	M/V	LOW					
Nitrate + Nitrite	Q2 Mean Factors	HIGH	M/V	HIGH	LOW	LOW					
pH	Q2 Mean Factors	M/V	M/V	M/V	M/V	LOW					
Total Phosphorus	Q2 Mean Factors	HIGH	HIGH	M/V	M/V	LOW					
Total Suspended Solids	Q2 Mean Factors	M/V	HIGH	M/V	LOW	M/V					
Specific Conductance	Q2 Mean Factors	HIGH	M/V	M/V	M/V	LOW					
Sulfate	Q2 Mean Factors	HIGH	M/V	LOW	M/V	LOW					
Temperature	Q2 Mean Factors	M/V	HIGH	LOW	M/V	M/V					
Turbidity	Q2 Mean Factors	LOW	HIGH	M/V	LOW	M/V					
Iron	Q2 Mean Factors	M/V	HIGH	M/V	LOW	M/V					

Supplementary Table 5.10 – The Quarter 2 Median Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Median Factors	LOW	HIGH	HIGH	LOW	M/V					
Total Organic Carbon	Q2 Median Factors	M/V	LOW	M/V	HIGH	M/V					
Chloride	Q2 Median Factors	LOW	M/V	HIGH	M/V	M/V					
Chemical Oxygen Demand	Q2 Median Factors	M/V	LOW	M/V	HIGH	M/V					
Dissolved Oxygen	Q2 Median Factors	LOW	M/V	M/V	M/V	HIGH					
Hardness	Q2 Median Factors	LOW	HIGH	HIGH	LOW	M/V					
Total Kjeldahl Nitrogen	Q2 Median Factors	LOW	LOW	M/V	HIGH	M/V					
Nitrate + Nitrite	Q2 Median Factors	LOW	HIGH	HIGH	M/V	LOW					
pH	Q2 Median Factors	LOW	HIGH	M/V	M/V	M/V					
Total Phosphorus	Q2 Median Factors	LOW	M/V	HIGH	HIGH	LOW					
Total Suspended Solids	Q2 Median Factors	M/V	M/V	M/V	HIGH	LOW					
Specific Conductance	Q2 Median Factors	LOW	M/V	HIGH	M/V	M/V					
Sulfate	Q2 Median Factors	LOW	LOW	HIGH	M/V	M/V					
Temperature	Q2 Median Factors	M/V	M/V	M/V	HIGH	M/V					
Turbidity	Q2 Median Factors	M/V	M/V	M/V	HIGH	LOW					
Iron	Q2 Median Factors	M/V	M/V	M/V	HIGH	LOW					

Supplementary Table 5.11 – The Quarter 2 Trimmed Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	M/V	HIGH	M/V			
Total Organic Carbon	Q2 Trimmed Mean Factors	M/V	HIGH	HIGH	M/V	M/V	LOW	M/V			
Chloride	Q2 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	HIGH	M/V	LOW			
Chemical Oxygen Demand	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	M/V	M/V	LOW	M/V			
Dissolved Oxygen	Q2 Trimmed Mean Factors	M/V	M/V	M/V	LOW	HIGH	HIGH	LOW			
Hardness	Q2 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	M/V	HIGH	LOW			
Total Kjeldahl Nitrogen	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	LOW	M/V	LOW	M/V			
Nitrate + Nitrite	Q2 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	M/V	HIGH	M/V			
pH	Q2 Trimmed Mean Factors	M/V	HIGH	M/V	LOW	M/V	HIGH	M/V			
Total Phosphorus	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	LOW	LOW	M/V	M/V			
Total Suspended Solids	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	M/V	LOW	M/V	HIGH			
Specific Conductance	Q2 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	M/V	M/V	LOW			
Sulfate	Q2 Trimmed Mean Factors	HIGH	LOW	HIGH	M/V	M/V	LOW	LOW			
Temperature	Q2 Trimmed Mean Factors	LOW	M/V	HIGH	M/V	M/V	M/V	M/V			
Turbidity	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	M/V	LOW	M/V	HIGH			
Iron	Q2 Trimmed Mean Factors	M/V	M/V	HIGH	M/V	LOW	M/V	HIGH			

Supplementary Table 5.12 – The Quarter 2 Geometric Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Geomean Factors	M/V	HIGH	LOW	LOW	M/V	LOW	HIGH			
Total Organic Carbon	Q2 Geomean Factors	M/V	M/V	M/V	HIGH	HIGH	M/V	LOW			
Chloride	Q2 Geomean Factors	HIGH	HIGH	LOW	M/V	M/V	LOW	M/V			
Chemical Oxygen Demand	Q2 Geomean Factors	M/V	M/V	M/V	M/V	HIGH	M/V	LOW			
Dissolved Oxygen	Q2 Geomean Factors	HIGH	M/V	LOW	M/V	M/V	LOW	HIGH			
Hardness	Q2 Geomean Factors	M/V	HIGH	LOW	LOW	M/V	LOW	HIGH			
Total Kjeldahl Nitrogen	Q2 Geomean Factors	M/V	M/V	M/V	M/V	HIGH	LOW	LOW			
Nitrate + Nitrite	Q2 Geomean Factors	M/V	HIGH	M/V	LOW	M/V	LOW	HIGH			
pH	Q2 Geomean Factors	M/V	M/V	M/V	HIGH	M/V	LOW	HIGH			
Total Phosphorus	Q2 Geomean Factors	M/V	M/V	M/V	M/V	HIGH	LOW	M/V			
Total Suspended Solids	Q2 Geomean Factors	LOW	M/V	HIGH	M/V	HIGH	M/V	M/V			
Specific Conductance	Q2 Geomean Factors	M/V	HIGH	LOW	LOW	M/V	LOW	M/V			
Sulfate	Q2 Geomean Factors	M/V	HIGH	LOW	LOW	HIGH	M/V	LOW			
Temperature	Q2 Geomean Factors	M/V	M/V	M/V	M/V	HIGH	M/V	M/V			
Turbidity	Q2 Geomean Factors	LOW	M/V	HIGH	M/V	HIGH	M/V	M/V			
Iron	Q2 Geomean Factors	LOW	M/V	HIGH	M/V	HIGH	M/V	LOW			

Supplementary Table 5.13 – The Quarter 3 Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Mean Factors	HIGH	HIGH	LOW	M/V	LOW	M/V	LOW			
Total Organic Carbon	Q3 Mean Factors	LOW	HIGH	M/V	LOW	HIGH	M/V	LOW			
Chloride	Q3 Mean Factors	M/V	HIGH	LOW	M/V	M/V	M/V	LOW			
Chemical Oxygen Demand	Q3 Mean Factors	LOW	M/V	M/V	LOW	M/V	HIGH	M/V			
Dissolved Oxygen	Q3 Mean Factors	M/V	M/V	LOW	M/V	M/V	HIGH	M/V			
Hardness	Q3 Mean Factors	HIGH	HIGH	LOW	M/V	LOW	M/V	LOW			
Total Kjeldahl Nitrogen	Q3 Mean Factors	LOW	M/V	LOW	LOW	M/V	HIGH	M/V			
Nitrate + Nitrite	Q3 Mean Factors	M/V	HIGH	M/V	M/V	LOW	M/V	LOW			
pH	Q3 Mean Factors	M/V	M/V	LOW	M/V	M/V	HIGH	M/V			
Total Phosphorus	Q3 Mean Factors	M/V	HIGH	LOW	LOW	M/V	HIGH	LOW			
Total Suspended Solids	Q3 Mean Factors	M/V	M/V	M/V	LOW	M/V	HIGH	HIGH			
Specific Conductance	Q3 Mean Factors	M/V	HIGH	LOW	M/V	M/V	M/V	LOW			
Sulfate	Q3 Mean Factors	LOW	HIGH	M/V	LOW	M/V	HIGH	LOW			
Temperature	Q3 Mean Factors	M/V	M/V	M/V	LOW	M/V	HIGH	M/V			
Turbidity	Q3 Mean Factors	M/V	M/V	M/V	LOW	LOW	HIGH	HIGH			
Iron	Q3 Mean Factors	M/V	M/V	HIGH	LOW	LOW	HIGH	HIGH			

Supplementary Table 5.14 – The Quarter 3 Median Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Median Factors	HIGH	HIGH	LOW	M/V	LOW					
Total Organic Carbon	Q3 Median Factors	HIGH	LOW	M/V	M/V	M/V					
Chloride	Q3 Median Factors	HIGH	M/V	M/V	M/V	LOW					
Chemical Oxygen Demand	Q3 Median Factors	M/V	LOW	HIGH	M/V	M/V					
Dissolved Oxygen	Q3 Median Factors	M/V	M/V	HIGH	HIGH	LOW					
Hardness	Q3 Median Factors	HIGH	HIGH	LOW	M/V	LOW					
Total Kjeldahl Nitrogen	Q3 Median Factors	M/V	LOW	HIGH	M/V	M/V					
Nitrate + Nitrite	Q3 Median Factors	HIGH	M/V	LOW	M/V	M/V					
pH	Q3 Median Factors	M/V	HIGH	HIGH	M/V	LOW					
Total Phosphorus	Q3 Median Factors	HIGH	M/V	M/V	LOW	M/V					
Total Suspended Solids	Q3 Median Factors	M/V	M/V	HIGH	LOW	M/V					
Specific Conductance	Q3 Median Factors	HIGH	M/V	M/V	M/V	LOW					
Sulfate	Q3 Median Factors	HIGH	LOW	M/V	LOW	M/V					
Temperature	Q3 Median Factors	M/V	LOW	HIGH	M/V	M/V					
Turbidity	Q3 Median Factors	M/V	M/V	HIGH	LOW	M/V					
Iron	Q3 Median Factors	M/V	M/V	HIGH	LOW	M/V					

Supplementary Table 5.15 – The Quarter 3 Trimmed Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Trimmed Mean Factors	LOW	LOW	HIGH	LOW	HIGH					
Total Organic Carbon	Q3 Trimmed Mean Factors	M/V	HIGH	LOW	M/V	HIGH					
Chloride	Q3 Trimmed Mean Factors	M/V	M/V	M/V	LOW	HIGH					
Chemical Oxygen Demand	Q3 Trimmed Mean Factors	HIGH	HIGH	LOW	M/V	M/V					
Dissolved Oxygen	Q3 Trimmed Mean Factors	HIGH	HIGH	M/V	LOW	M/V					
Hardness	Q3 Trimmed Mean Factors	LOW	LOW	HIGH	LOW	HIGH					
Total Kjeldahl Nitrogen	Q3 Trimmed Mean Factors	HIGH	M/V	LOW	LOW	M/V					
Nitrate + Nitrite	Q3 Trimmed Mean Factors	LOW	LOW	M/V	M/V	HIGH					
pH	Q3 Trimmed Mean Factors	HIGH	M/V	M/V	LOW	M/V					
Total Phosphorus	Q3 Trimmed Mean Factors	M/V	M/V	M/V	M/V	HIGH					
Total Suspended Solids	Q3 Trimmed Mean Factors	HIGH	M/V	M/V	M/V	M/V					
Specific Conductance	Q3 Trimmed Mean Factors	M/V	M/V	M/V	LOW	HIGH					
Sulfate	Q3 Trimmed Mean Factors	M/V	LOW	LOW	M/V	HIGH					
Temperature	Q3 Trimmed Mean Factors	HIGH	M/V	LOW	M/V	M/V					
Turbidity	Q3 Trimmed Mean Factors	HIGH	LOW	M/V	M/V	M/V					
Iron	Q3 Trimmed Mean Factors	HIGH	LOW	M/V	HIGH	M/V					

Supplementary Table 5.16 – The Quarter 3 Geometric Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Geomean Factors	LOW	HIGH	LOW	HIGH	M/V					
Total Organic Carbon	Q3 Geomean Factors	M/V	M/V	M/V	LOW	M/V					
Chloride	Q3 Geomean Factors	M/V	HIGH	LOW	M/V	M/V					
Chemical Oxygen Demand	Q3 Geomean Factors	HIGH	M/V	M/V	LOW	M/V					
Dissolved Oxygen	Q3 Geomean Factors	HIGH	M/V	LOW	M/V	HIGH					
Hardness	Q3 Geomean Factors	LOW	HIGH	LOW	M/V	M/V					
Total Kjeldahl Nitrogen	Q3 Geomean Factors	HIGH	M/V	M/V	LOW	M/V					
Nitrate + Nitrite	Q3 Geomean Factors	LOW	HIGH	M/V	M/V	LOW					
pH	Q3 Geomean Factors	HIGH	M/V	LOW	M/V	M/V					
Total Phosphorus	Q3 Geomean Factors	M/V	HIGH	M/V	M/V	LOW					
Total Suspended Solids	Q3 Geomean Factors	HIGH	M/V	M/V	M/V	LOW					
Specific Conductance	Q3 Geomean Factors	M/V	HIGH	LOW	M/V	M/V					
Sulfate	Q3 Geomean Factors	M/V	HIGH	M/V	LOW	LOW					
Temperature	Q3 Geomean Factors	HIGH	M/V	M/V	LOW	M/V					
Turbidity	Q3 Geomean Factors	HIGH	M/V	M/V	M/V	LOW					
Iron	Q3 Geomean Factors	HIGH	M/V	M/V	M/V	LOW					

Supplementary Table 5.17 – The Quarter 4 Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Mean Factors	HIGH	LOW	LOW	LOW	HIGH					
Total Organic Carbon	Q4 Mean Factors	LOW	M/V	M/V	HIGH	HIGH					
Chloride	Q4 Mean Factors	M/V	M/V	LOW	M/V	HIGH					
Chemical Oxygen Demand	Q4 Mean Factors	LOW	HIGH	M/V	M/V	M/V					
Dissolved Oxygen	Q4 Mean Factors	M/V	M/V	LOW	HIGH	M/V					
Hardness	Q4 Mean Factors	HIGH	LOW	LOW	LOW	HIGH					
Total Kjeldahl Nitrogen	Q4 Mean Factors	LOW	HIGH	LOW	M/V	HIGH					
Nitrate + Nitrite	Q4 Mean Factors	HIGH	M/V	LOW	LOW	HIGH					
pH	Q4 Mean Factors	M/V	HIGH	LOW	M/V	M/V					
Total Phosphorus	Q4 Mean Factors	M/V	M/V	M/V	M/V	HIGH					
Total Suspended Solids	Q4 Mean Factors	LOW	HIGH	M/V	M/V	M/V					
Specific Conductance	Q4 Mean Factors	M/V	M/V	LOW	M/V	HIGH					
Sulfate	Q4 Mean Factors	M/V	M/V	M/V	LOW	HIGH					
Temperature	Q4 Mean Factors	LOW	M/V	M/V	M/V	M/V					
Turbidity	Q4 Mean Factors	LOW	HIGH	M/V	M/V	M/V					
Iron	Q4 Mean Factors	LOW	HIGH	M/V	M/V	M/V					

Supplementary Table 5.18 – The Quarter 4 Median Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Median Factors	LOW	HIGH	M/V	M/V	M/V	M/V	HIGH			
Total Organic Carbon	Q4 Median Factors	M/V	LOW	HIGH	HIGH	LOW	HIGH	M/V			
Chloride	Q4 Median Factors	LOW	M/V	M/V	M/V	LOW	HIGH	HIGH			
Chemical Oxygen Demand	Q4 Median Factors	M/V	LOW	HIGH	HIGH	M/V	HIGH	M/V			
Dissolved Oxygen	Q4 Median Factors	LOW	HIGH	M/V	HIGH	M/V	LOW	M/V			
Hardness	Q4 Median Factors	LOW	HIGH	M/V	M/V	M/V	M/V	HIGH			
Total Kjeldahl Nitrogen	Q4 Median Factors	M/V	LOW	HIGH	M/V	M/V	HIGH	M/V			
Nitrate + Nitrite	Q4 Median Factors	LOW	M/V	M/V	LOW	M/V	M/V	HIGH			
pH	Q4 Median Factors	LOW	HIGH	HIGH	M/V	M/V	M/V	M/V			
Total Phosphorus	Q4 Median Factors	M/V	M/V	HIGH	M/V	LOW	HIGH	M/V			
Total Suspended Solids	Q4 Median Factors	M/V	LOW	HIGH	M/V	HIGH	M/V	LOW			
Specific Conductance	Q4 Median Factors	LOW	M/V	M/V	M/V	LOW	HIGH	HIGH			
Sulfate	Q4 Median Factors	M/V	LOW	HIGH	LOW	LOW	HIGH	HIGH			
Temperature	Q4 Median Factors	M/V	M/V	HIGH	M/V	M/V	HIGH	LOW			
Turbidity	Q4 Median Factors	M/V	LOW	HIGH	M/V	HIGH	M/V	LOW			
Iron	Q4 Median Factors	M/V	LOW	HIGH	LOW	HIGH	M/V	LOW			

Supplementary Table 5.19 – The Quarter 4 Trimmed Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Trimmed Mean Factors	HIGH	HIGH	M/V	LOW	LOW	M/V				
Total Organic Carbon	Q4 Trimmed Mean Factors	LOW	M/V	M/V	M/V	HIGH	HIGH				
Chloride	Q4 Trimmed Mean Factors	M/V	HIGH	LOW	LOW	M/V	M/V				
Chemical Oxygen Demand	Q4 Trimmed Mean Factors	LOW	M/V	M/V	M/V	M/V	HIGH				
Dissolved Oxygen	Q4 Trimmed Mean Factors	HIGH	M/V	M/V	LOW	HIGH	M/V				
Hardness	Q4 Trimmed Mean Factors	HIGH	HIGH	M/V	LOW	LOW	M/V				
Total Kjeldahl Nitrogen	Q4 Trimmed Mean Factors	LOW	HIGH	M/V	LOW	M/V	HIGH				
Nitrate + Nitrite	Q4 Trimmed Mean Factors	M/V	HIGH	M/V	LOW	LOW	M/V				
pH	Q4 Trimmed Mean Factors	HIGH	LOW	M/V	LOW	M/V	HIGH				
Total Phosphorus	Q4 Trimmed Mean Factors	M/V	HIGH	LOW	M/V	M/V	HIGH				
Total Suspended Solids	Q4 Trimmed Mean Factors	LOW	M/V	HIGH	M/V	LOW	HIGH				
Specific Conductance	Q4 Trimmed Mean Factors	M/V	HIGH	LOW	LOW	M/V	M/V				
Sulfate	Q4 Trimmed Mean Factors	M/V	HIGH	LOW	M/V	LOW	HIGH				
Temperature	Q4 Trimmed Mean Factors	M/V	M/V	M/V	M/V	M/V	HIGH				
Turbidity	Q4 Trimmed Mean Factors	LOW	M/V	HIGH	M/V	M/V	HIGH				
Iron	Q4 Trimmed Mean Factors	LOW	M/V	HIGH	M/V	M/V	HIGH				

Supplementary Table 5.20 – The Quarter 4 Geometric Mean Factor clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Geomean Factors	M/V	M/V	HIGH	LOW	HIGH	LOW	M/V			
Total Organic Carbon	Q4 Geomean Factors	HIGH	M/V	M/V	M/V	LOW	HIGH	HIGH			
Chloride	Q4 Geomean Factors	M/V	LOW	HIGH	LOW	M/V	LOW	M/V			
Chemical Oxygen Demand	Q4 Geomean Factors	M/V	M/V	M/V	M/V	LOW	M/V	HIGH			
Dissolved Oxygen	Q4 Geomean Factors	M/V	M/V	M/V	LOW	HIGH	HIGH	M/V			
Hardness	Q4 Geomean Factors	M/V	M/V	HIGH	LOW	HIGH	LOW	M/V			
Total Kjeldahl Nitrogen	Q4 Geomean Factors	M/V	M/V	HIGH	LOW	LOW	M/V	HIGH			
Nitrate + Nitrite	Q4 Geomean Factors	LOW	M/V	HIGH	LOW	M/V	M/V	M/V			
pH	Q4 Geomean Factors	M/V	M/V	M/V	LOW	HIGH	M/V	HIGH			
Total Phosphorus	Q4 Geomean Factors	M/V	LOW	HIGH	M/V	M/V	M/V	HIGH			
Total Suspended Solids	Q4 Geomean Factors	M/V	HIGH	M/V	M/V	M/V	M/V	HIGH			
Specific Conductance	Q4 Geomean Factors	M/V	LOW	HIGH	LOW	M/V	LOW	M/V			
Sulfate	Q4 Geomean Factors	M/V	LOW	HIGH	M/V	LOW	LOW	HIGH			
Temperature	Q4 Geomean Factors	M/V	M/V	M/V	M/V	M/V	M/V	HIGH			
Turbidity	Q4 Geomean Factors	M/V	HIGH	M/V	M/V	LOW	M/V	HIGH			
Iron	Q4 Geomean Factors	M/V	HIGH	M/V	M/V	LOW	M/V	HIGH			

Supplementary Table 5.21 – The Annual Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Mean SOM	HIGH	LOW	M/V							
Total Organic Carbon	Annual Mean SOM	M/V	M/V	HIGH							
Chloride	Annual Mean SOM	M/V	LOW	HIGH							
Chemical Oxygen Demand	Annual Mean SOM	LOW	M/V	HIGH							
Dissolved Oxygen	Annual Mean SOM	M/V	LOW	M/V							
Hardness	Annual Mean SOM	HIGH	LOW	M/V							
Total Kjeldahl Nitrogen	Annual Mean SOM	M/V	LOW	HIGH							
Nitrate + Nitrite	Annual Mean SOM	M/V	LOW	M/V							
pH	Annual Mean SOM	M/V	M/V	M/V							
Total Phosphorus	Annual Mean SOM	M/V	M/V	HIGH							
Total Suspended Solids	Annual Mean SOM	LOW	M/V	HIGH							
Specific Conductance	Annual Mean SOM	HIGH	LOW	HIGH							
Sulfate	Annual Mean SOM	M/V	LOW	HIGH							
Temperature	Annual Mean SOM	LOW	M/V	HIGH							
Turbidity	Annual Mean SOM	LOW	M/V	HIGH							
Iron	Annual Mean SOM	LOW	HIGH	HIGH							

Supplementary Table 5.22 – The Annual Median SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Median SOM	M/V	HIGH	LOW	HIGH	LOW	M/V	LOW	HIGH		
Total Organic Carbon	Annual Median SOM	HIGH	LOW	M/V	M/V	HIGH	M/V	LOW	HIGH		
Chloride	Annual Median SOM	M/V	M/V	LOW	HIGH	LOW	M/V	LOW	HIGH		
Chemical Oxygen Demand	Annual Median SOM	HIGH	LOW	M/V	M/V	M/V	HIGH	M/V	HIGH		
Dissolved Oxygen	Annual Median SOM	M/V	M/V	LOW	M/V	M/V	HIGH	M/V	M/V		
Hardness	Annual Median SOM	M/V	HIGH	LOW	HIGH	LOW	M/V	LOW	M/V		
Total Kjeldahl Nitrogen	Annual Median SOM	HIGH	LOW	LOW	M/V	M/V	M/V	HIGH	HIGH		
Nitrate + Nitrite	Annual Median SOM	M/V	M/V	LOW	HIGH	M/V	M/V	LOW	M/V		
pH	Annual Median SOM	M/V	HIGH	LOW	M/V	M/V	M/V	LOW	M/V		
Total Phosphorus	Annual Median SOM	HIGH	LOW	M/V	HIGH	M/V	M/V	M/V	HIGH		
Total Suspended Solids	Annual Median SOM	HIGH	M/V	M/V	LOW	M/V	M/V	HIGH	M/V		
Specific Conductance	Annual Median SOM	M/V	M/V	LOW	HIGH	LOW	M/V	LOW	HIGH		
Sulfate	Annual Median SOM	M/V	LOW	M/V	HIGH	LOW	M/V	M/V	HIGH		
Temperature	Annual Median SOM	HIGH	M/V	M/V	M/V	M/V	M/V	M/V	HIGH		
Turbidity	Annual Median SOM	HIGH	M/V	M/V	M/V	M/V	M/V	HIGH	M/V		
Iron	Annual Median SOM	HIGH	LOW	M/V	M/V	M/V	M/V	HIGH	M/V		

Supplementary Table 5.23 – The Annual Trimmed Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Trimmed Mean SOM	M/V	LOW	M/V	HIGH	M/V	HIGH	LOW			
Total Organic Carbon	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	M/V	M/V			
Chloride	Annual Trimmed Mean SOM	HIGH	LOW	M/V	M/V	M/V	HIGH	LOW			
Chemical Oxygen Demand	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	M/V	M/V			
Dissolved Oxygen	Annual Trimmed Mean SOM	M/V	M/V	HIGH	M/V	M/V	M/V	LOW			
Hardness	Annual Trimmed Mean SOM	M/V	LOW	M/V	HIGH	M/V	HIGH	LOW			
Total Kjeldahl Nitrogen	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	M/V	LOW			
Nitrate + Nitrite	Annual Trimmed Mean SOM	M/V	LOW	M/V	M/V	M/V	HIGH	LOW			
pH	Annual Trimmed Mean SOM	M/V	M/V	M/V	M/V	HIGH	M/V	LOW			
Total Phosphorus	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	HIGH	LOW			
Total Suspended Solids	Annual Trimmed Mean SOM	M/V	HIGH	M/V	M/V	HIGH	M/V	M/V			
Specific Conductance	Annual Trimmed Mean SOM	HIGH	LOW	M/V	M/V	M/V	HIGH	LOW			
Sulfate	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	M/V	HIGH	LOW			
Temperature	Annual Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	LOW	M/V			
Turbidity	Annual Trimmed Mean SOM	M/V	HIGH	LOW	M/V	HIGH	M/V	M/V			
Iron	Annual Trimmed Mean SOM	M/V	HIGH	LOW	M/V	HIGH	LOW	M/V			

Supplementary Table 5.24 – The Annual Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Annual Geomean SOM	LOW	LOW	M/V	HIGH	HIGH	M/V	M/V	LOW		
Total Organic Carbon	Annual Geomean SOM	M/V	HIGH	M/V	M/V	LOW	HIGH	HIGH	M/V		
Chloride	Annual Geomean SOM	LOW	LOW	M/V	HIGH	M/V	M/V	HIGH	LOW		
Chemical Oxygen Demand	Annual Geomean SOM	M/V	M/V	M/V	M/V	LOW	HIGH	HIGH	M/V		
Dissolved Oxygen	Annual Geomean SOM	M/V	M/V	HIGH	M/V	M/V	M/V	M/V	LOW		
Hardness	Annual Geomean SOM	LOW	LOW	M/V	HIGH	HIGH	M/V	M/V	LOW		
Total Kjeldahl Nitrogen	Annual Geomean SOM	M/V	M/V	M/V	M/V	LOW	HIGH	HIGH	LOW		
Nitrate + Nitrite	Annual Geomean SOM	LOW	M/V	M/V	HIGH	M/V	M/V	M/V	LOW		
pH	Annual Geomean SOM	M/V	M/V	LOW	M/V	M/V	HIGH	M/V	LOW		
Total Phosphorus	Annual Geomean SOM	M/V	M/V	M/V	HIGH	LOW	HIGH	HIGH	M/V		
Total Suspended Solids	Annual Geomean SOM	HIGH	M/V	M/V	M/V	M/V	HIGH	M/V	M/V		
Specific Conductance	Annual Geomean SOM	LOW	LOW	M/V	HIGH	M/V	M/V	HIGH	LOW		
Sulfate	Annual Geomean SOM	M/V	LOW	M/V	HIGH	LOW	HIGH	HIGH	M/V		
Temperature	Annual Geomean SOM	M/V	M/V	M/V	LOW	LOW	HIGH	HIGH	M/V		
Turbidity	Annual Geomean SOM	HIGH	M/V	M/V	M/V	M/V	HIGH	M/V	M/V		
Iron	Annual Geomean SOM	HIGH	M/V	M/V	LOW	LOW	HIGH	M/V	M/V		

Supplementary Table 5.25 – The Quarter 1 Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Mean SOM	M/V	LOW	LOW	M/V	HIGH	M/V	HIGH			
Total Organic Carbon	Q1 Mean SOM	M/V	M/V	M/V	HIGH	HIGH	M/V	LOW			
Chloride	Q1 Mean SOM	M/V	LOW	LOW	HIGH	HIGH	M/V	M/V			
Chemical Oxygen Demand	Q1 Mean SOM	HIGH	M/V	M/V	HIGH	HIGH	M/V	LOW			
Dissolved Oxygen	Q1 Mean SOM	M/V	M/V	M/V	LOW	M/V	M/V	M/V			
Hardness	Q1 Mean SOM	M/V	LOW	LOW	M/V	M/V	M/V	HIGH			
Total Kjeldahl Nitrogen	Q1 Mean SOM	HIGH	M/V	LOW	HIGH	M/V	M/V	LOW			
Nitrate + Nitrite	Q1 Mean SOM	LOW	M/V	LOW	M/V	M/V	HIGH	HIGH			
pH	Q1 Mean SOM	M/V	M/V	LOW	LOW	M/V	M/V	HIGH			
Total Phosphorus	Q1 Mean SOM	HIGH	HIGH	LOW	HIGH	HIGH	M/V	LOW			
Total Suspended Solids	Q1 Mean SOM	HIGH	HIGH	M/V	HIGH	M/V	M/V	LOW			
Specific Conductance	Q1 Mean SOM	M/V	LOW	LOW	HIGH	HIGH	M/V	M/V			
Sulfate	Q1 Mean SOM	M/V	M/V	LOW	HIGH	HIGH	M/V	M/V			
Temperature	Q1 Mean SOM	HIGH	M/V	M/V	HIGH	LOW	M/V	M/V			
Turbidity	Q1 Mean SOM	HIGH	HIGH	M/V	M/V	M/V	HIGH	LOW			
Iron	Q1 Mean SOM	HIGH	HIGH	M/V	M/V	M/V	HIGH	LOW			

Supplementary Table 5.26 – The Quarter 1 Median SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Median SOM	M/V	HIGH	HIGH	M/V	LOW	LOW				
Total Organic Carbon	Q1 Median SOM	HIGH	M/V	LOW	HIGH	HIGH	M/V				
Chloride	Q1 Median SOM	HIGH	HIGH	LOW	HIGH	M/V	LOW				
Chemical Oxygen Demand	Q1 Median SOM	HIGH	M/V	LOW	M/V	HIGH	M/V				
Dissolved Oxygen	Q1 Median SOM	LOW	M/V	M/V	M/V	M/V	M/V				
Hardness	Q1 Median SOM	M/V	HIGH	HIGH	M/V	LOW	LOW				
Total Kjeldahl Nitrogen	Q1 Median SOM	HIGH	M/V	LOW	M/V	M/V	LOW				
Nitrate + Nitrite	Q1 Median SOM	M/V	M/V	HIGH	LOW	LOW	LOW				
pH	Q1 Median SOM	M/V	M/V	M/V	HIGH	M/V	LOW				
Total Phosphorus	Q1 Median SOM	HIGH	M/V	LOW	M/V	HIGH	LOW				
Total Suspended Solids	Q1 Median SOM	M/V	M/V	LOW	M/V	HIGH	M/V				
Specific Conductance	Q1 Median SOM	HIGH	HIGH	M/V	M/V	M/V	LOW				
Sulfate	Q1 Median SOM	HIGH	HIGH	LOW	M/V	M/V	LOW				
Temperature	Q1 Median SOM	HIGH	LOW	M/V	M/V	M/V	M/V				
Turbidity	Q1 Median SOM	HIGH	M/V	LOW	M/V	HIGH	M/V				
Iron	Q1 Median SOM	M/V	M/V	LOW	LOW	HIGH	M/V				

Supplementary Table 5.27 – The Quarter 1 Trimmed Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	LOW				
Total Organic Carbon	Q1 Trimmed Mean SOM	M/V	HIGH	HIGH	M/V	LOW	M/V				
Chloride	Q1 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW	M/V	LOW				
Chemical Oxygen Demand	Q1 Trimmed Mean SOM	M/V	HIGH	HIGH	HIGH	LOW	M/V				
Dissolved Oxygen	Q1 Trimmed Mean SOM	M/V	LOW	LOW	M/V	M/V	M/V				
Hardness	Q1 Trimmed Mean SOM	HIGH	M/V	M/V	LOW	HIGH	LOW				
Total Kjeldahl Nitrogen	Q1 Trimmed Mean SOM	HIGH	HIGH	HIGH	M/V	LOW	LOW				
Nitrate + Nitrite	Q1 Trimmed Mean SOM	M/V	M/V	M/V	M/V	HIGH	LOW				
pH	Q1 Trimmed Mean SOM	M/V	LOW	M/V	LOW	HIGH	M/V				
Total Phosphorus	Q1 Trimmed Mean SOM	M/V	HIGH	HIGH	M/V	LOW	LOW				
Total Suspended Solids	Q1 Trimmed Mean SOM	M/V	M/V	HIGH	HIGH	LOW	M/V				
Specific Conductance	Q1 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW	M/V	LOW				
Sulfate	Q1 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW	M/V	LOW				
Temperature	Q1 Trimmed Mean SOM	M/V	HIGH	HIGH	M/V	LOW	M/V				
Turbidity	Q1 Trimmed Mean SOM	M/V	M/V	HIGH	HIGH	LOW	M/V				
Iron	Q1 Trimmed Mean SOM	LOW	M/V	HIGH	HIGH	LOW	M/V				

Supplementary Table 5.28 – The Quarter 1 Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q1 Geomean SOM	HIGH	LOW	LOW	M/V	LOW	M/V	HIGH	M/V	M/V	
Total Organic Carbon	Q1 Geomean SOM	LOW	M/V	LOW	M/V	M/V	HIGH	M/V	HIGH	HIGH	
Chloride	Q1 Geomean SOM	M/V	LOW	LOW	M/V	LOW	HIGH	HIGH	M/V	HIGH	
Chemical Oxygen Demand	Q1 Geomean SOM	LOW	M/V	LOW	M/V	M/V	M/V	M/V	HIGH	HIGH	
Dissolved Oxygen	Q1 Geomean SOM	HIGH	HIGH	M/V	LOW	M/V	M/V	M/V	M/V	LOW	
Hardness	Q1 Geomean SOM	HIGH	LOW	LOW	M/V	LOW	M/V	HIGH	M/V	M/V	
Total Kjeldahl Nitrogen	Q1 Geomean SOM	LOW	M/V	LOW	M/V	M/V	M/V	M/V	HIGH	HIGH	
Nitrate + Nitrite	Q1 Geomean SOM	HIGH	M/V	LOW	M/V	M/V	M/V	M/V	M/V	M/V	
pH	Q1 Geomean SOM	M/V	HIGH	LOW	M/V	LOW	M/V	M/V	M/V	LOW	
Total Phosphorus	Q1 Geomean SOM	LOW	M/V	LOW	M/V	M/V	M/V	M/V	HIGH	HIGH	
Total Suspended Solids	Q1 Geomean SOM	LOW	M/V	M/V	M/V	HIGH	M/V	M/V	HIGH	M/V	
Specific Conductance	Q1 Geomean SOM	M/V	LOW	LOW	M/V	LOW	M/V	HIGH	M/V	HIGH	
Sulfate	Q1 Geomean SOM	LOW	LOW	M/V	M/V	LOW	M/V	HIGH	M/V	HIGH	
Temperature	Q1 Geomean SOM	LOW	M/V	M/V	LOW	M/V	M/V	M/V	HIGH	HIGH	
Turbidity	Q1 Geomean SOM	LOW	M/V	M/V	M/V	HIGH	M/V	M/V	HIGH	M/V	
Iron	Q1 Geomean SOM	LOW	M/V	M/V	M/V	HIGH	M/V	LOW	HIGH	M/V	

Supplementary Table 5.29 – The Quarter 2 Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Mean SOM	LOW	M/V	HIGH	LOW	M/V	HIGH	LOW	HIGH	M/V	
Total Organic Carbon	Q2 Mean SOM	LOW	M/V	M/V	HIGH	M/V	M/V	M/V	LOW	HIGH	
Chloride	Q2 Mean SOM	LOW	HIGH	HIGH	LOW	M/V	M/V	LOW	M/V	M/V	
Chemical Oxygen Demand	Q2 Mean SOM	LOW	M/V	M/V	M/V	LOW	LOW	M/V	LOW	HIGH	
Dissolved Oxygen	Q2 Mean SOM	LOW	HIGH	M/V	M/V	M/V	M/V	M/V	M/V	M/V	
Hardness	Q2 Mean SOM	LOW	M/V	HIGH	LOW	M/V	HIGH	LOW	HIGH	M/V	
Total Kjeldahl Nitrogen	Q2 Mean SOM	LOW	M/V	HIGH	M/V	M/V	M/V	M/V	LOW	HIGH	
Nitrate + Nitrite	Q2 Mean SOM	LOW	LOW	HIGH	LOW	HIGH	HIGH	M/V	HIGH	M/V	
pH	Q2 Mean SOM	LOW	M/V	M/V	M/V	M/V	M/V	M/V	M/V	M/V	
Total Phosphorus	Q2 Mean SOM	LOW	M/V	HIGH	M/V	M/V	M/V	M/V	M/V	HIGH	
Total Suspended Solids	Q2 Mean SOM	M/V	LOW	M/V	M/V	HIGH	M/V	HIGH	LOW	HIGH	
Specific Conductance	Q2 Mean SOM	LOW	M/V	HIGH	LOW	M/V	M/V	LOW	M/V	M/V	
Sulfate	Q2 Mean SOM	M/V	M/V	HIGH	LOW	M/V	M/V	M/V	LOW	HIGH	
Temperature	Q2 Mean SOM	M/V	M/V	M/V	M/V	M/V	LOW	M/V	LOW	HIGH	
Turbidity	Q2 Mean SOM	LOW	LOW	M/V	M/V	HIGH	M/V	HIGH	M/V	HIGH	
Iron	Q2 Mean SOM	M/V	LOW	M/V	M/V	HIGH	M/V	HIGH	M/V	HIGH	

Supplementary Table 5.30 – The Quarter 2 Median SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Median SOM	LOW	HIGH	LOW	HIGH	LOW	M/V	HIGH	M/V		
Total Organic Carbon	Q2 Median SOM	LOW	LOW	HIGH	LOW	M/V	M/V	HIGH	HIGH		
Chloride	Q2 Median SOM	LOW	LOW	LOW	HIGH	LOW	HIGH	HIGH	M/V		
Chemical Oxygen Demand	Q2 Median SOM	LOW	LOW	M/V	LOW	HIGH	M/V	M/V	HIGH		
Dissolved Oxygen	Q2 Median SOM	LOW	M/V	M/V	M/V	LOW	HIGH	M/V	M/V		
Hardness	Q2 Median SOM	LOW	HIGH	LOW	HIGH	LOW	M/V	HIGH	M/V		
Total Kjeldahl Nitrogen	Q2 Median SOM	LOW	LOW	M/V	LOW	M/V	M/V	HIGH	HIGH		
Nitrate + Nitrite	Q2 Median SOM	LOW	HIGH	LOW	HIGH	M/V	M/V	HIGH	M/V		
pH	Q2 Median SOM	LOW	M/V	M/V	M/V	LOW	M/V	M/V	M/V		
Total Phosphorus	Q2 Median SOM	LOW	LOW	M/V	M/V	M/V	LOW	HIGH	HIGH		
Total Suspended Solids	Q2 Median SOM	M/V	M/V	M/V	M/V	HIGH	M/V	M/V	HIGH		
Specific Conductance	Q2 Median SOM	LOW	M/V	LOW	HIGH	LOW	M/V	HIGH	M/V		
Sulfate	Q2 Median SOM	M/V	LOW	LOW	M/V	LOW	M/V	HIGH	HIGH		
Temperature	Q2 Median SOM	M/V	M/V	M/V	LOW	M/V	M/V	M/V	HIGH		
Turbidity	Q2 Median SOM	M/V	M/V	M/V	M/V	HIGH	LOW	M/V	HIGH		
Iron	Q2 Median SOM	M/V	M/V	M/V	LOW	HIGH	LOW	M/V	HIGH		

Supplementary Table 5.31 – The Quarter 2 Trimmed Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Trimmed Mean SOM	LOW	HIGH	LOW	M/V	HIGH	M/V				
Total Organic Carbon	Q2 Trimmed Mean SOM	LOW	LOW	HIGH	M/V	M/V	HIGH				
Chloride	Q2 Trimmed Mean SOM	LOW	M/V	LOW	HIGH	HIGH	M/V				
Chemical Oxygen Demand	Q2 Trimmed Mean SOM	LOW	LOW	M/V	M/V	M/V	HIGH				
Dissolved Oxygen	Q2 Trimmed Mean SOM	LOW	M/V	LOW	HIGH	M/V	M/V				
Hardness	Q2 Trimmed Mean SOM	LOW	HIGH	LOW	M/V	HIGH	M/V				
Total Kjeldahl Nitrogen	Q2 Trimmed Mean SOM	LOW	LOW	M/V	M/V	HIGH	HIGH				
Nitrate + Nitrite	Q2 Trimmed Mean SOM	LOW	HIGH	LOW	LOW	HIGH	M/V				
pH	Q2 Trimmed Mean SOM	LOW	M/V	M/V	M/V	M/V	M/V				
Total Phosphorus	Q2 Trimmed Mean SOM	LOW	M/V	M/V	M/V	HIGH	HIGH				
Total Suspended Solids	Q2 Trimmed Mean SOM	M/V	M/V	M/V	LOW	M/V	HIGH				
Specific Conductance	Q2 Trimmed Mean SOM	LOW	M/V	LOW	M/V	HIGH	M/V				
Sulfate	Q2 Trimmed Mean SOM	M/V	M/V	LOW	M/V	HIGH	M/V				
Temperature	Q2 Trimmed Mean SOM	M/V	LOW	M/V	M/V	M/V	HIGH				
Turbidity	Q2 Trimmed Mean SOM	M/V	M/V	HIGH	LOW	M/V	HIGH				
Iron	Q2 Trimmed Mean SOM	M/V	M/V	M/V	LOW	M/V	HIGH				

Supplementary Table 5.32 – The Quarter 2 Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q2 Geomean SOM	LOW	M/V	HIGH	LOW	LOW	M/V	HIGH	M/V	HIGH	
Total Organic Carbon	Q2 Geomean SOM	LOW	M/V	LOW	HIGH	M/V	M/V	HIGH	HIGH	HIGH	
Chloride	Q2 Geomean SOM	LOW	LOW	M/V	LOW	LOW	HIGH	M/V	M/V	HIGH	
Chemical Oxygen Demand	Q2 Geomean SOM	LOW	M/V	LOW	M/V	M/V	M/V	M/V	HIGH	HIGH	
Dissolved Oxygen	Q2 Geomean SOM	LOW	M/V	M/V	M/V	M/V	HIGH	M/V	M/V	M/V	
Hardness	Q2 Geomean SOM	LOW	M/V	HIGH	LOW	LOW	M/V	HIGH	M/V	M/V	
Total Kjeldahl Nitrogen	Q2 Geomean SOM	LOW	LOW	LOW	M/V	M/V	M/V	HIGH	HIGH	HIGH	
Nitrate + Nitrite	Q2 Geomean SOM	LOW	M/V	HIGH	LOW	M/V	M/V	M/V	M/V	M/V	
pH	Q2 Geomean SOM	LOW	M/V	M/V	M/V	M/V	M/V	M/V	M/V	M/V	
Total Phosphorus	Q2 Geomean SOM	LOW	M/V	LOW	M/V	M/V	M/V	HIGH	HIGH	HIGH	
Total Suspended Solids	Q2 Geomean SOM	M/V	M/V	LOW	M/V	HIGH	M/V	M/V	HIGH	M/V	
Specific Conductance	Q2 Geomean SOM	LOW	M/V	HIGH	LOW	LOW	M/V	HIGH	M/V	HIGH	
Sulfate	Q2 Geomean SOM	M/V	M/V	M/V	LOW	M/V	M/V	HIGH	HIGH	HIGH	
Temperature	Q2 Geomean SOM	M/V	M/V	LOW	M/V	M/V	M/V	M/V	HIGH	M/V	
Turbidity	Q2 Geomean SOM	M/V	M/V	LOW	M/V	HIGH	LOW	M/V	HIGH	M/V	
Iron	Q2 Geomean SOM	M/V	M/V	LOW	M/V	HIGH	LOW	M/V	HIGH	M/V	

Supplementary Table 5.33 – The Quarter 3 Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Mean SOM	HIGH	LOW	HIGH	M/V						
Total Organic Carbon	Q3 Mean SOM	LOW	M/V	HIGH	M/V						
Chloride	Q3 Mean SOM	M/V	LOW	HIGH	M/V						
Chemical Oxygen Demand	Q3 Mean SOM	LOW	M/V	M/V	HIGH						
Dissolved Oxygen	Q3 Mean SOM	M/V	M/V	M/V	HIGH						
Hardness	Q3 Mean SOM	HIGH	LOW	HIGH	M/V						
Total Kjeldahl Nitrogen	Q3 Mean SOM	LOW	M/V	HIGH	HIGH						
Nitrate + Nitrite	Q3 Mean SOM	M/V	LOW	HIGH	M/V						
pH	Q3 Mean SOM	M/V	LOW	M/V	HIGH						
Total Phosphorus	Q3 Mean SOM	M/V	LOW	HIGH	HIGH						
Total Suspended Solids	Q3 Mean SOM	LOW	M/V	M/V	HIGH						
Specific Conductance	Q3 Mean SOM	M/V	LOW	HIGH	M/V						
Sulfate	Q3 Mean SOM	M/V	LOW	HIGH	HIGH						
Temperature	Q3 Mean SOM	LOW	M/V	M/V	HIGH						
Turbidity	Q3 Mean SOM	M/V	M/V	M/V	HIGH						
Iron	Q3 Mean SOM	M/V	M/V	M/V	HIGH						

Supplementary Table 5.34 – The Quarter 3 Median SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Median SOM	HIGH	LOW	HIGH	M/V	LOW					
Total Organic Carbon	Q3 Median SOM	LOW	M/V	M/V	HIGH	M/V					
Chloride	Q3 Median SOM	M/V	LOW	HIGH	HIGH	M/V					
Chemical Oxygen Demand	Q3 Median SOM	LOW	M/V	M/V	HIGH	HIGH					
Dissolved Oxygen	Q3 Median SOM	M/V	M/V	M/V	HIGH	HIGH					
Hardness	Q3 Median SOM	HIGH	LOW	HIGH	HIGH	M/V					
Total Kjeldahl Nitrogen	Q3 Median SOM	LOW	M/V	M/V	HIGH	HIGH					
Nitrate + Nitrite	Q3 Median SOM	M/V	LOW	HIGH	HIGH	M/V					
pH	Q3 Median SOM	HIGH	LOW	M/V	M/V	HIGH					
Total Phosphorus	Q3 Median SOM	M/V	LOW	HIGH	HIGH	M/V					
Total Suspended Solids	Q3 Median SOM	M/V	M/V	M/V	M/V	HIGH					
Specific Conductance	Q3 Median SOM	M/V	LOW	HIGH	HIGH	M/V					
Sulfate	Q3 Median SOM	LOW	LOW	HIGH	HIGH	M/V					
Temperature	Q3 Median SOM	LOW	M/V	LOW	M/V	HIGH					
Turbidity	Q3 Median SOM	M/V	M/V	M/V	M/V	HIGH					
Iron	Q3 Median SOM	M/V	M/V	M/V	M/V	HIGH					

Supplementary Table 5.35 – The Quarter 3 Trimmed Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	HIGH						
Total Organic Carbon	Q3 Trimmed Mean SOM	HIGH	M/V	M/V	LOW						
Chloride	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	M/V						
Chemical Oxygen Demand	Q3 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW						
Dissolved Oxygen	Q3 Trimmed Mean SOM	M/V	HIGH	M/V	M/V						
Hardness	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	HIGH						
Total Kjeldahl Nitrogen	Q3 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW						
Nitrate + Nitrite	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	M/V						
pH	Q3 Trimmed Mean SOM	M/V	HIGH	LOW	M/V						
Total Phosphorus	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	M/V						
Total Suspended Solids	Q3 Trimmed Mean SOM	M/V	HIGH	M/V	LOW						
Specific Conductance	Q3 Trimmed Mean SOM	HIGH	M/V	LOW	M/V						
Sulfate	Q3 Trimmed Mean SOM	HIGH	HIGH	LOW	M/V						
Temperature	Q3 Trimmed Mean SOM	M/V	HIGH	M/V	LOW						
Turbidity	Q3 Trimmed Mean SOM	M/V	HIGH	M/V	M/V						
Iron	Q3 Trimmed Mean SOM	M/V	HIGH	M/V	M/V						

Supplementary Table 5.36 – The Quarter 3 Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q3 Geomean SOM	M/V	HIGH	HIGH	M/V	M/V	LOW	HIGH			
Total Organic Carbon	Q3 Geomean SOM	M/V	HIGH	M/V	M/V	M/V	M/V	LOW			
Chloride	Q3 Geomean SOM	M/V	HIGH	HIGH	M/V	M/V	LOW	M/V			
Chemical Oxygen Demand	Q3 Geomean SOM	HIGH	HIGH	M/V	HIGH	M/V	M/V	LOW			
Dissolved Oxygen	Q3 Geomean SOM	HIGH	M/V	M/V	M/V	HIGH	LOW	M/V			
Hardness	Q3 Geomean SOM	M/V	HIGH	HIGH	LOW	M/V	LOW	HIGH			
Total Kjeldahl Nitrogen	Q3 Geomean SOM	HIGH	HIGH	M/V	HIGH	M/V	M/V	LOW			
Nitrate + Nitrite	Q3 Geomean SOM	M/V	HIGH	HIGH	M/V	M/V	LOW	M/V			
pH	Q3 Geomean SOM	HIGH	M/V	M/V	M/V	M/V	LOW	M/V			
Total Phosphorus	Q3 Geomean SOM	M/V	HIGH	HIGH	M/V	M/V	LOW	M/V			
Total Suspended Solids	Q3 Geomean SOM	HIGH	HIGH	M/V	HIGH	M/V	M/V	M/V			
Specific Conductance	Q3 Geomean SOM	M/V	HIGH	HIGH	M/V	M/V	LOW	M/V			
Sulfate	Q3 Geomean SOM	HIGH	HIGH	HIGH	LOW	M/V	LOW	LOW			
Temperature	Q3 Geomean SOM	HIGH	HIGH	LOW	M/V	M/V	M/V	LOW			
Turbidity	Q3 Geomean SOM	HIGH	M/V	M/V	HIGH	M/V	M/V	M/V			
Iron	Q3 Geomean SOM	HIGH	M/V	M/V	HIGH	M/V	M/V	M/V			

Supplementary Table 5.37 – The Quarter 4 Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Mean SOM	M/V	HIGH	HIGH	LOW	LOW					
Total Organic Carbon	Q4 Mean SOM	HIGH	M/V	LOW	M/V	M/V					
Chloride	Q4 Mean SOM	HIGH	HIGH	M/V	LOW	M/V					
Chemical Oxygen Demand	Q4 Mean SOM	HIGH	M/V	LOW	M/V	HIGH					
Dissolved Oxygen	Q4 Mean SOM	M/V	M/V	M/V	M/V	M/V					
Hardness	Q4 Mean SOM	M/V	HIGH	HIGH	LOW	M/V					
Total Kjeldahl Nitrogen	Q4 Mean SOM	HIGH	M/V	LOW	M/V	HIGH					
Nitrate + Nitrite	Q4 Mean SOM	HIGH	HIGH	M/V	LOW	LOW					
pH	Q4 Mean SOM	M/V	M/V	M/V	LOW	M/V					
Total Phosphorus	Q4 Mean SOM	HIGH	M/V	LOW	LOW	M/V					
Total Suspended Solids	Q4 Mean SOM	HIGH	LOW	LOW	M/V	HIGH					
Specific Conductance	Q4 Mean SOM	HIGH	HIGH	M/V	LOW	M/V					
Sulfate	Q4 Mean SOM	HIGH	HIGH	LOW	LOW	M/V					
Temperature	Q4 Mean SOM	HIGH	M/V	LOW	M/V	M/V					
Turbidity	Q4 Mean SOM	M/V	M/V	LOW	M/V	HIGH					
Iron	Q4 Mean SOM	M/V	LOW	M/V	M/V	HIGH					

Supplementary Table 5.38 – The Quarter 4 Median SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Median SOM	M/V	M/V	LOW	LOW	HIGH	HIGH				
Total Organic Carbon	Q4 Median SOM	HIGH	M/V	M/V	HIGH	M/V	LOW				
Chloride	Q4 Median SOM	HIGH	LOW	LOW	M/V	HIGH	M/V				
Chemical Oxygen Demand	Q4 Median SOM	HIGH	M/V	M/V	HIGH	M/V	LOW				
Dissolved Oxygen	Q4 Median SOM	M/V	M/V	LOW	HIGH	M/V	M/V				
Hardness	Q4 Median SOM	M/V	LOW	LOW	LOW	HIGH	HIGH				
Total Kjeldahl Nitrogen	Q4 Median SOM	HIGH	M/V	M/V	M/V	M/V	LOW				
Nitrate + Nitrite	Q4 Median SOM	M/V	LOW	LOW	LOW	HIGH	M/V				
pH	Q4 Median SOM	M/V	M/V	LOW	M/V	M/V	M/V				
Total Phosphorus	Q4 Median SOM	HIGH	LOW	M/V	M/V	HIGH	LOW				
Total Suspended Solids	Q4 Median SOM	HIGH	HIGH	M/V	M/V	LOW	LOW				
Specific Conductance	Q4 Median SOM	HIGH	LOW	LOW	LOW	HIGH	M/V				
Sulfate	Q4 Median SOM	HIGH	M/V	M/V	LOW	HIGH	M/V				
Temperature	Q4 Median SOM	HIGH	M/V	M/V	M/V	M/V	M/V				
Turbidity	Q4 Median SOM	HIGH	HIGH	M/V	M/V	M/V	LOW				
Iron	Q4 Median SOM	HIGH	HIGH	M/V	M/V	M/V	LOW				

Supplementary Table 5.39 – The Quarter 4 Trimmed Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Trimmed Mean SOM	HIGH	HIGH	M/V	LOW	M/V	LOW	LOW			
Total Organic Carbon	Q4 Trimmed Mean SOM	M/V	LOW	HIGH	M/V	M/V	M/V	M/V			
Chloride	Q4 Trimmed Mean SOM	HIGH	M/V	HIGH	M/V	M/V	LOW	LOW			
Chemical Oxygen Demand	Q4 Trimmed Mean SOM	M/V	LOW	HIGH	M/V	HIGH	M/V	M/V			
Dissolved Oxygen	Q4 Trimmed Mean SOM	M/V	HIGH	M/V	M/V	M/V	M/V	LOW			
Hardness	Q4 Trimmed Mean SOM	HIGH	HIGH	HIGH	LOW	M/V	LOW	LOW			
Total Kjeldahl Nitrogen	Q4 Trimmed Mean SOM	M/V	LOW	HIGH	M/V	HIGH	M/V	LOW			
Nitrate + Nitrite	Q4 Trimmed Mean SOM	HIGH	M/V	M/V	LOW	M/V	LOW	LOW			
pH	Q4 Trimmed Mean SOM	M/V	M/V	M/V	M/V	M/V	M/V	LOW			
Total Phosphorus	Q4 Trimmed Mean SOM	HIGH	M/V	HIGH	M/V	M/V	LOW	M/V			
Total Suspended Solids	Q4 Trimmed Mean SOM	LOW	LOW	HIGH	M/V	HIGH	HIGH	M/V			
Specific Conductance	Q4 Trimmed Mean SOM	HIGH	M/V	HIGH	LOW	M/V	LOW	LOW			
Sulfate	Q4 Trimmed Mean SOM	HIGH	M/V	HIGH	LOW	M/V	M/V	M/V			
Temperature	Q4 Trimmed Mean SOM	M/V	LOW	HIGH	M/V	HIGH	M/V	M/V			
Turbidity	Q4 Trimmed Mean SOM	M/V	LOW	M/V	M/V	HIGH	HIGH	M/V			
Iron	Q4 Trimmed Mean SOM	LOW	LOW	M/V	M/V	HIGH	HIGH	M/V			

Supplementary Table 5.40 – The Quarter 4 Geometric Mean SOM clusters' T-test results for determining if the mean of each variable in a given cluster was significantly different (α=0.05) than the mean of that variable for all of the stations

Variable	Dataset Clustered	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Alkalinity	Q4 Geomean SOM	HIGH	M/V	M/V	LOW	HIGH					
Total Organic Carbon	Q4 Geomean SOM	M/V	HIGH	M/V	M/V	LOW					
Chloride	Q4 Geomean SOM	HIGH	HIGH	M/V	LOW	M/V					
Chemical Oxygen Demand	Q4 Geomean SOM	M/V	HIGH	HIGH	M/V	LOW					
Dissolved Oxygen	Q4 Geomean SOM	M/V	M/V	M/V	M/V	M/V					
Hardness	Q4 Geomean SOM	HIGH	M/V	M/V	LOW	HIGH					
Total Kjeldahl Nitrogen	Q4 Geomean SOM	M/V	HIGH	HIGH	M/V	LOW					
Nitrate + Nitrite	Q4 Geomean SOM	HIGH	HIGH	M/V	LOW	M/V					
pH	Q4 Geomean SOM	M/V	M/V	M/V	LOW	M/V					
Total Phosphorus	Q4 Geomean SOM	HIGH	HIGH	M/V	LOW	LOW					
Total Suspended Solids	Q4 Geomean SOM	LOW	M/V	HIGH	M/V	LOW					
Specific Conductance	Q4 Geomean SOM	HIGH	HIGH	M/V	LOW	M/V					
Sulfate	Q4 Geomean SOM	HIGH	HIGH	M/V	LOW	LOW					
Temperature	Q4 Geomean SOM	M/V	HIGH	HIGH	M/V	LOW					
Turbidity	Q4 Geomean SOM	LOW	M/V	HIGH	M/V	LOW					
Iron	Q4 Geomean SOM	LOW	M/V	HIGH	M/V	LOW					

Cluster Consistency

Supplementary Table 6.1 – Stations that Clustered Consistently among Different Statistical Indicators in the Annual Datasets

Annual Factor Consistent Clusters	Annual SOM ConsistentClusters
BL-.7, EW-239, FR-17	BL-.7, EW-239, FC-26, FR-17, FR-64, MC-35, SGR-1, WR-319, WR-348
BWC-4, SGR-1	BL-64, CIC-17, EC-21, WLC-2
EC-1, FC-0.6, FC-7	BWC-4, EW-168, MC-18
EC-21, WR-192, WR-210	EC-7, FC-0.6, FC-7
EEL-1, EEL-38, EW-79, MC-18	EEL-1, EEL-38, EW-1, EW-94
EW-1, EW-94	GC-8, VF-38
FC-26, FR-64, MC-35, WLC-2	IN-2, LST-2, MU-20, SLT-12
IN-2, SLT-12	IWC-9, WR-248, WR-293, WR-309
IWC-9, WR-248	WR-134, WR-162
WR-19, WR-81	WR-19, WR-46, WR-81
WR-293, WR-309	WR-192, WR-210
15 STATIONS WERE VARIABLE	4 STATIONS WERE VARIABLE

Supplementary Table 6.2 – Stations that Clustered Consistently among Different Statistical Indicators in the Quarter 1 Datasets

Quarter 1 Factor Consistent Clusters	Quarter 1 SOM Consistent Clusters
BL-.7, SGR-1, WR-248	BL-.7, EW-239, FC-26, FR-17, FR-64, MC-35, SGR-1, WR-348
BL-64, FC-7	BWC-4, CIC-17, EW-168, MC-18
CIC-17, MC-18	EC-1, EC-7, FC-0.6
EEL-1, EEL-38, EW-1, EW-79, EW-94, WR-134, WR-19, WR-46, WR-81	EEL-1, EEL-38, EW-1
FC-0.6, IWC-9, WR-248, WR-279, WR-293, WR-309	EW-79, MU-20
GC-8, VF-38	GC-8, VF-38
IN-2, LST-2	IN-2, LST-2, SLT-12
MU-20, SLT-12	IWC-9, WLC-2, WR-279, WR-293, WR-309
WR-192, WR-210	WR-162, WR-192, WR-210
	WR-19, WR-46, WR-81
14 STATIONS WERE VARIABLE	8 STATIONS WERE VARIABLE

Supplementary Table 6.3 – Stations that Clustered Consistently among Different Statistical Indicators in the Quarter 2 Datasets

Quarter 2 Factor Consistent Clusters	Quarter 2 SOM Consistent Clusters
BL-.7, BWC-4, EW-239, FR-17, FR-64, SGR-1, WR-348	BL-64, CIC-17, EC-21, FC-26
BL-64, EC-21, IWC-9, WR-210, WR-248, WR-279, WR-293, WR-309, WR-319	BWC-4, EW-168, MC-18, WR-319
CIC-17, FC-26	EC-1, EC-7, FC-0.6
EC-1, EC-7, FC-0.6, FC-7	EEL-38, EW-1, EW-94
EEL-1, EEL-38, EW-1, EW-94	EW-239, FR-17, FR-64, MC-35, WR-348
EW-168, MC-18, SND-4	GC-8, VF-38
GC-8, VF-38	IN-2, LST-2, SLT-12
IN-2, MU-20, SLT-12	IWC-9, WR-248, WR-279, WR-293, WR-309
WR-134, WR-162, WR-19, WR-46, WR-81	WR-134, WR-162, WR-19, WR-46, WR-81
	WR-192, WR-210
5 STATIONS WERE VARIABLE	8 STATIONS WERE VARIABLE

Supplementary Table 6.4 – Stations that Clustered Consistently among Different Statistical Indicators in the Quarter 3 Datasets

Quarter 3 Factor Consistent Clusters	Quarter 3 SOM Consistent Clusters
BL-.7, BWC-4, EW-168, EW-239, FC-26, MC-18, WLC-2	BL-.7, BWC-4, EW-168, EW-239, FC-26, FR-17, FR-64, MC-18, MC-35, SGR-1, WLC-2, WR-319, WR-348
BL-64, EC-21, IWC-9, WR-192, WR-210, WR-248, WR-279, WR-293, WR-309	CIC-17, EC-21, WR-293
EC-1, EC-7, GC-8, VF-38, EEL-1, EW-1, EW-79, EW-94	EC-7, FC-7
FC-0.6, FC-7	EEL-1, EW-94
FR-17, FR-64, SGR-1	EW-79, GC-8, IN-2, LST-2, MU-20, SLT-12, SND-4, VF-38
IN-2, LST-2, MU-20, SLT-12	IWC-9, WR-279, WR-309
WR-134, WR-162, WR-19, WR-46, WR-81	WR-134, WR-19, WR-46, WR-81
	WR-192, WR-210
6 STATIONS WERE VARIABLE	7 STATIONS WERE VARIABLE

Supplementary Table 6.5 – Stations that Clustered Consistently among Different Statistical Indicators in the Quarter 4 Datasets

Quarter 4 Factor Consistent Clusters	Quarter 4 SOM Consistent Clusters
BL-.7, EW-239, FC-26, FR-17, MC-35, SGR-1, WLC-2	BL-.7, BWC-4, EW-168, EW-239, FC-26, FR-17, FR-64, MC-35, SGR-1, WR-319, WR-348
EC-1, EC-7, FC-0.6, FC-7	BL-64, WLC-2
EC-21, IWC-9, WR-248, WR-309	EC-21, IWC-9, WR-248, WR-279, WR-293
EEL-1, EEL-38, EW-1	EC-7, FC-0.6, FC-7, GC-8, SND-4, VF-38
EW-79, EW-94, MC-18	EEL-1, EEL-38, EW-1
GC-8, VF-38	EW-79, EW-94
IN-2, LST-2, MU-20, SLT-12	IN-2, LST-2, MU-20, SLT-12
WR-134, WR-19, WR-46, WR-81	WR-134, WR-162, WR-192, WR-210
WR-192, WR-210	WR-19, WR-46, WR-81
WR-279, WR-293	
9 STATIONS WERE VARIABLE	4 STATIONS WERE VARIABLE

Supplementary Table 6.6 – Stations that Clustered Consistently among Different Quarters and the Annual Dataset when the Mean was the Statistical Indicator

Mean Factor Consistent Clusters	Mean SOM Consistent Clusters
BL-.7, WR-348	BL-.7, EW-239, FR-17, FR-64, MC-35, WR-348
BWC-4, EW-239	BL-64, FC-26
EC-1, EC-7, VF-38, GC-8	BWC-4, EW-168, MC-18
EEL-1, EEL-38, EW-1	EC-7, FC-0.6
FR-17, SGR-1	EEL-1, EEL-38, EW-1
FR-64, MC-35	EW-79, GC-8, MU-20, VF-38
IN-2, LST-2	IN-2, LST-2, SLT-12
IWC-9, WR-348, WR-309	IWC-9, WR-248, WR-293
MU-20, SLT-12	SGR-1, WR-319
WR-134, WR-19, WR-46, WR-81	WR-19, WR-46, WR-81
WR-192, WR-210	WR-192, WR-210
16 STATIONS WERE VARIABLE	11 STATIONS WERE VARIABLE

Supplementary Table 6.7 – Stations that Clustered Consistently among Different Quarterly and the Annual Datasets when the Median was the Statistical Indicator

Median Factor Consistent Clusters	**Median SOM Consistent Clusters**
BL-.7, EW-239, SGR-1, WLC-2	BL-.7, FC-26, SGR-1
BWC-4, FC-26, FR-17, FR-64	BL-64, EC-21
EC-1, EC-7	BWC-4, EW-168, EW-239, FR-17, FR-64, MC-18, MC-35, WR-319, WR-348
EC-21, IWC-9, WR-248, WR-293, WR-309	EC-7, FC-0.6, FC-7
EEL-1, EW-1, EW-94	EW-79, EW-94
GC-8, VF-38	GC-8, VF-38
IN-2, LST-2	IN-2, LST-2, SLT-12
MU-20, SLT-12	IWC-9, WR-279, WR-293, WR-309
WR-134, WR-162, WR-19, WR-46, WR-81	WR-19, WR-46, WR-81
WR-192, WR-210	WR-192, WR-210
13 STATIONS WERE VARIABLE	11 STATIONS WERE VARIABLE

Supplementary Table 6.8 – Stations that Clustered Consistently among Different Quarterly and the Annual Datasets when the Trimmed Mean was the Statistical Indicator

Trimmed Mean Factor Consistent Clusters	**Trimmed Mean SOM Consistent Clusters**
BL-.7, EW-239, FR-17	BL-.7, BWC.4, EW-168, EW-239, FC-26, FR-17, FR-64, MC-35, SGR-1, WR-319, WR-348
EEL-1, EEL-38, EW-79	CIC-17, EC-21
EW-1, EW-94	EEL-38, EW-1, EW-94
IWC-9, WR-248	GC-8, VF-38
WR-19, WR-81	IN-2, LST-2, SLT-12
	IWC-9, WR-248, WR-279, WR-309
	WR-19, WR-46, WR-81
	WR-192, WR-210
32 STATIONS WERE VARIABLE	14 STATIONS WERE VARIABLE

Supplementary Table 6.9 – Stations that Clustered Consistently among Different Quarterly and the Annual Datasets when the Geometric Mean was the Statistical Indicator

Geometric Mean Factor Consistent Clusters	Geometric Mean SOM Consistent Clusters
BL-.7, BWC-4, EW-239, FR-17, FR-64, SGR-1, WLC-2	BL-.7, EW-239, FC-26, FR-17, FR-64, MC-35, SGR-1, WR-348
BL-64, EC-21, IWC-9, WR-192, WR-210, WR-248, WR-279, WR-293, WR-309	BL-64, EC-21
EC-1, FC-0.6, FC-7	BWC-4, EW-168, MC-18
EEL-1, EEL-38, EW-1, EW-79, EW-94	EC-7, FC-0.6
GC-8, VF-38	EEL-1, EW-1, EW-94
IN-2, MU-20, SLT-12	EEL-38, EW-79
WR-134, WR-19, WR-46, WR-81	GC-8, VF-38
	IN-2, LST-2, SLT-12
	IWC-9, WR-279, WR-293, WR-309
	WR-19, WR-46, WR-81
	WR-192, WR-210
11 STATIONS ARE VARIABLE	10 STATIONS ARE VARIABLE

Supplementary Table 6.10 – Stations that Clustered Consistently for the Factor Based Cluster Assignments and the SOM Based Cluster Assignments

Overall Factor Consistent Clusters	Overall SOM Consistent Clusters
IWC-9, WR-248	BWC-4, EW-168
WR-19, WR-81	EW-239, FR-17, FR-64, MC-35, WR-248
	GC-8, VF-38
	IN-2, LST-2, SLT-12
	WR-19, WR-46, WR-81
	WR-192, WR-210
40 STATIONS WERE VARIABLE	27 STATIONS WERE VARIABLE

LDA Classification Equations

Supplementary Table 7.1 – Classification coefficients and constant for the annual mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Constant	-299.2	-241.5	-264.5	-414.9	-265.5	-252.7
Interior Plateau	120.3	152.3	113.5	316.1	112.3	111.5
Drainage Area	-233.6	-174.7	-242.3	-342.4	-204.7	-154.9
Cultivated Crops	389.8	324.1	300.4	371.9	432.7	328.1
Shawnee Hills	34.7	-2.5	2.7	-61.6	69.3	21.8
Moderately Well Drained Soil	47.3	68.3	48.0	49.0	18.0	71.6
CAFO	-8.7	4.2	2.2	-19.0	-30.9	3.2
Temperature	472.7	391.6	494.3	597.5	400.2	388.8
Forest to Urban	8.9	21.8	-61.0	-46.6	70.2	21.2
Water	147.1	123.4	187.3	203.4	91.5	119.0
Agriculture to Urban	122.9	119.7	163.4	159.9	87.9	117.5
Highland Rim	38.3	-5.3	38.9	42.8	5.2	8.8

Supplementary Table 7.2 – Classification coefficients and constant for the annual median LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-586.9	-464.4	-601.5	-622.5	-555.2
Highland Rim	-34.6	-0.7	-38.6	-8.3	-0.4
NPDES	164.9	105.4	150.1	189.8	124.2
Cultivated Crops	889.1	723.9	928.8	887.1	846.7
Slope %	648.5	510.2	677.6	616.7	625.2
Forest to Urban	264.7	85.0	298.7	245.5	221.0
Water	-103.5	15.8	-130.2	-87.8	-77.5
Precipitation	-107.1	30.7	-81.6	-88.6	-12.6
Agriculture to Urban	-66.1	53.1	-75.4	-66.9	-38.8
Grassland, Pasture, Scrubland	255.8	170.1	254.4	251.6	220.1
Wetlands	222.4	147.8	231.8	210.6	214.3
CAFO	12.9	1.7	-11.2	12.3	-8.7
Agriculture to Forest	223.6	240.5	194.9	236.4	195.2
SSLimestoneShl	-55.3	-68.0	-58.5	-24.6	-61.1

Supplementary Table 7.3 – Classification coefficients and constant for the annual trimmed mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-23.6	-20.8	-21.8	-21.8	-17.3
Agriculture to Urban	-7.7	26.1	-13.6	5.3	-3.3
Forest to Urban	38.3	7.9	47.0	35.8	34.1
Cultivated Crops	40.3	24.1	31.4	33.7	30.0
Moderately Well Drained Soil	10.5	3.9	25.3	10.7	15.1
Sum of Streams	4.3	8.6	-8.8	-2.7	-1.0

Supplementary Table 7.4 – Classification coefficients and constant for the annual geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-350.6	-245.8	-289.4	-281.3	-225.1
Highland Rim	139.4	102.7	132.0	102.9	101.2
NPDES	29.0	4.2	-22.5	-27.5	-2.0
Precipitation	236.3	215.4	263.8	269.5	200.2
Urban	170.7	155.8	159.7	170.0	121.7
Agriculture to Urban	-160.9	-134.7	-129.0	-118.8	-101.4
Wetlands	150.2	141.7	160.2	163.1	121.2
Urban to Agriculture	-41.0	-35.7	-47.3	-48.9	-25.2
Sandstone, Limestone, Shale	284.3	224.8	230.6	219.3	223.3
Eastern Corn Belt	391.6	335.6	344.4	345.0	327.9

Supplementary Table 7.5 – Classification coefficients and constant for the quarter 1 mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-303.7	-593.8	-277.3	-569.4	-133.8
Forest	165.5	70.9	258.8	334.1	141.7
Interior Plateau	500.8	747.5	291.4	661.6	271.4
Longest Flow Path	149.9	110.9	-7.9	94.1	14.5
Temperature	59.0	633.6	110.9	253.7	40.3
Sandstone, Limestone, Shale	-219.9	-108.3	-51.4	-218.0	-88.3
NPDES	278.7	323.6	174.5	336.7	147.4
Poorly Drained Soil	201.4	157.4	204.0	254.3	149.9
Drainage Area	-273.2	-514.7	-93.7	-350.7	-96.7
Urban to Agriculture	-119.9	-171.3	-50.4	-160.5	-42.3
Gray Shale	165.0	43.5	129.8	182.1	102.9
Water	96.9	-40.6	75.4	84.2	69.9
ND	112.0	-206.4	108.0	40.3	82.5
Grassland, Pasture, Scrubland	-124.6	92.5	-154.1	-120.5	-99.2

Supplementary Table 7.5 (cont) – Classification coefficients and constant for the quarter 1 mean LDA classification equations

Variable	Cluster6	Cluster7	Cluster8	Cluster9	Cluster10
Constant	-143.0	-242.7	-457.5	-213.1	-131.6
Forest	127.8	87.6	680.1	163.4	127.4
Interior Plateau	198.5	283.8	210.9	314.2	234.9
Longest Flow Path	-68.0	-71.9	-73.1	-68.1	-61.1
Temperature	43.2	-24.8	-22.5	33.7	16.4
Sandstone, Limestone, Shale	-19.8	-68.0	-109.0	-61.6	-60.8
NPDES	106.7	145.0	220.7	135.8	123.2
Poorly Drained Soil	158.5	189.0	250.8	207.8	151.2
Drainage Area	-9.8	-19.1	-26.1	-17.9	-13.8
Urban to Agriculture	8.4	-46.4	-33.6	-18.9	-22.9
Gray Shale	101.8	216.0	139.8	126.8	113.3
Water	71.0	153.1	63.4	79.2	79.3
ND	123.2	188.7	121.8	132.9	107.1
Grassland, Pasture, Scrubland	-141.8	-186.5	-219.5	-160.3	-114.3

Supplementary Table 7.6 – Classification coefficients and constant for the quarter 1 median LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-1117.0	-1217.0	-1079.0	-968.6	-1142.0
Forest	1705.0	1873.0	1655.0	1628.0	1829.0
Interior Plateau	-306.0	-383.6	-257.3	-275.7	-356.1
Longest Flow Path	444.6	489.7	433.8	419.2	494.3
Water	307.1	293.2	314.8	297.2	268.2
Forest to Urban	419.6	445.9	440.5	386.3	432.5
Network Density	717.1	777.4	677.6	627.6	831.3
Cultivated Crops	1628.0	1663.0	1601.0	1517.0	1598.0
Agriculture to Forest	-24.0	40.0	-14.6	-30.7	41.3
Shawnee Hills	-113.4	-90.7	-131.3	-118.5	-114.3
Poorly Drained Soil	54.2	-3.4	60.5	45.7	-24.3
CAFO	-87.2	-70.1	-111.9	-86.3	-35.5
CSO	346.0	379.2	327.2	313.9	402.8
Temperature	-732.5	-884.7	-682.5	-671.6	-915.3
Grassland, Pasture, Scrubland	-252.2	-326.8	-246.4	-218.6	-353.7
Bluegrss	-37.3	-2.8	-45.8	-47.4	-1.2
Gray Shale	372.7	409.7	369.2	356.9	403.0
Highland Rim	-26.9	0.6	-46.5	-27.6	-46.3

Supplementary Table 7.6 (cont) – Classification coefficients and constant for the quarter 1 median LDA classification equations

Variable	Cluster6	Cluster7	Cluster8	Cluster9
Constant	-1235.0	-818.3	-1302.0	-1118.0
Forest	2259.0	571.9	1776.0	1671.0
Interior Plateau	-426.3	475.5	-53.4	-327.8
Longest Flow Path	553.4	83.4	351.4	381.2
Water	232.7	209.3	334.6	294.4
Forest to Urban	404.8	258.6	497.7	503.5
Network Density	905.8	-24.4	507.7	669.2
Cultivated Crops	1535.0	1131.0	1739.0	1676.0
Agriculture to Forest	156.0	-95.0	-143.4	-12.6
Shawnee Hills	-199.0	-201.7	-221.8	-60.7
Poorly Drained Soil	-63.9	198.4	164.6	58.5
CAFO	14.2	-235.6	-219.3	-146.6
CSO	431.0	76.1	264.9	304.2
Temperature	-1300.0	583.3	-362.0	-696.8
Grassland, Pasture, Scrubland	-467.3	205.7	-131.0	-266.4
Bluegrass	51.2	-317.5	-151.9	0.8
Gray Shale	434.5	183.7	365.5	368.6
Highland Rim	-27.8	-115.1	-71.1	-21.1

Supplementary Table 7.7 – Classification coefficients and constant for the quarter 1 trimmed mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-1060.0	-1009.0	-1118.0	-1141.0
Temperature	523.6	395.0	574.0	922.2
Interior Plateau	-226.7	-210.0	-18.8	226.9
Cultivated Crops	800.1	753.2	702.4	589.6
Shawnee Hills	125.9	54.3	-4.2	76.5
Forest	1654.0	1776.0	1691.0	1067.0
Forest to Urban	117.5	101.4	115.4	0.6
Agriculture to Forest	-71.7	-83.7	-136.7	58.5
Central Till Plain	1294.0	1246.0	1268.0	1276.0
Gray Shale	98.0	125.7	136.5	107.0
Moderately Well Drained Soil	-138.3	-132.2	-97.3	-113.3
NPDES	152.5	173.3	216.1	243.8
Drainage Area	-266.8	-249.9	-345.5	-474.7

Supplementary Table 7.7 (cont.) – Classification coefficients and constant for the quarter 1 trimmed mean LDA classification equations

Variable	Cluster5	Cluster6	Cluster7	Cluster8
Constant	-945.1	-948.6	-1004.0	-915.0
Temperature	456.9	470.6	437.4	495.0
Interior Plateau	-231.6	-154.7	-244.2	-183.7
Cultivated Crops	724.9	696.6	769.4	742.2
Shawnee Hills	75.3	48.6	93.6	112.9
Forest	1658.0	1620.0	1717.0	1504.0
Forest to Urban	112.0	102.0	111.8	119.9
Agriculture to Forest	-102.6	-97.0	-86.3	-73.0
Central Till Plain	1257.0	1232.0	1283.0	1141.0
Gray Shale	89.1	113.4	96.5	108.0
Moderately Well Drained Soil	-122.0	-116.6	-141.5	-105.2
NPDES	152.7	196.0	149.9	144.7
Drainage Area	-236.8	-274.6	-239.3	-235.6

Supplementary Table 7.8 – Classification coefficients and constant for the quarter 1 geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-686.0	-727.7	-890.8	-845.1	-828.4
Interior Plateau	1052.0	1109.0	1303.0	1168.0	1176.0
Drainage Area	-188.7	-243.0	-293.4	-275.4	-267.2
Wetlands	279.5	265.3	346.8	325.8	320.8
Forest	270.0	247.8	368.4	343.3	294.5
Shawnee Hills	142.1	161.2	38.5	124.7	137.7
Grassland, Pasture, Scrubland	91.4	85.1	130.1	110.1	106.9
Eastern Corn Belt	1198.0	1285.0	1279.0	1336.0	1338.0
NPDES	54.5	77.9	81.3	87.1	87.6

Supplementary Table 7.9 – Classification coefficients and constant for the quarter 2 mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-593.5	-492.5	-616.0	-426.4	-516.1
Cultivated Crops	931.2	827.0	980.7	773.0	832.2
Interior Plateau	772.3	659.8	811.6	651.5	769.2
Sandstone, Limestone, Shale	-260.9	-187.3	-265.2	-219.9	-252.3
Wetlands	35.6	42.5	6.4	13.9	40.3
Highland Rim	-254.7	-193.4	-266.7	-214.7	-217.6
Agriculture to Urban	22.5	-3.8	22.3	64.3	5.3
NPDES	230.8	207.2	238.0	180.1	199.7
Forest to Agriculture	291.5	261.3	307.1	256.4	270.4
Grassland, Pasture, Scrubland	252.9	222.7	241.3	209.8	230.4
Limestone	126.2	102.1	113.8	113.1	104.1
Forest to Urban	136.6	129.1	134.4	100.5	127.8

Supplementary Table 7.10 – Classification coefficients and constant for the quarter 2 median LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-172.2	-157.4	-171.7	-176.8	-127.7
Interior Plateau	316.3	230.3	286.6	227.6	213.6
Longest Flow Path	117.8	76.6	127.8	122.2	93.2
Cultivated Crops	155.4	204.8	178.1	172.6	154.0
Moderately Well Drained Soil	108.7	80.0	132.7	109.2	76.1
Sandstone, Limestone, Shale	-157.8	-92.6	-170.1	-98.0	-115.0
Highland Rim	-102.3	-99.3	-133.3	-86.9	-100.6
Grassland, Pasture, Scrubland	117.7	68.2	126.6	103.1	82.7
Network Density	9.7	49.4	11.3	22.5	57.7
Forest to Urban	67.6	88.6	71.9	72.1	32.7
Water	-8.0	-19.3	-6.5	-8.0	7.6
Agriculture to Urban	18.2	38.2	33.2	24.8	57.1

Supplementary Table 7.11 – Classification coefficients and constant for the quarter 2 trimmed mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-5260.0	-4027.0	-4429.0	-2646.0
Forest	9152.0	8130.0	8356.0	6241.0
Drainage Area	6374.0	5205.0	5746.0	3868.0
Interior Plateau	-1496.0	-1185.0	-1334.0	-553.7
Shawnee Hills	-211.6	-350.7	-215.8	-357.2
Wetlands	2766.0	2359.0	2522.0	1842.0
Longest Flow Path	-4517.0	-3792.0	-4089.0	-2896.0
NPDES	368.8	441.9	432.4	298.9
Urban	1553.0	1206.0	1365.0	961.6
Highland Rim	-1124.0	-866.3	-964.4	-670.0
CAFOs	3762.0	3216.0	3438.0	2363.0
Central Till Plain	5238.0	4484.0	4770.0	3478.0
CSOs	1218.0	945.7	1069.0	768.6
Agriculture to Forest	712.6	741.2	720.7	439.8
Precipitation	-3078.0	-2617.0	-2805.0	-1764.0
Urban to Forest	-819.5	-685.7	-764.2	-570.4
Gray Shale	-249.7	-262.1	-238.5	-82.3
Poorly Drained Soil	-870.4	-681.2	-771.0	-443.4

Supplementary Table 7.11(cont.) – Classification coefficients and constant for the quarter 2 trimmed mean LDA classification equations

Variable	Cluster5	Cluster6	Cluster7
Constant	-5201.0	-3619.0	-3318.0
Forest	8982.0	7569.0	7279.0
Drainage Area	6389.0	5185.0	4846.0
Interior Plateau	-1432.0	-1176.0	-1081.0
Shawnee Hills	-171.3	-195.3	-232.6
Wetlands	2752.0	2252.0	2160.0
Longest Flow Path	-4560.0	-3682.0	-3474.0
NPDES	272.5	299.5	327.8
Urban	1636.0	1277.0	1177.0
Highland Rim	-1153.0	-921.2	-816.3
CAFOs	3639.0	3053.0	2932.0
Central Till Plain	5172.0	4373.0	4126.0
CSOs	1237.0	987.4	925.2
Agriculture to Forest	568.0	560.9	578.4
Precipitation	-2892.0	-2443.0	-2343.0
Urban to Forest	-810.5	-675.3	-650.6
Gray Shale	-185.8	-187.8	-186.0
Poorly Drained Soil	-818.4	-677.3	-641.4

Supplementary Table 7.12 – Classification coefficients and constant for the quarter 2 geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-646.8	-555.8	-587.2	-1166.0
Forest	1036.0	1272.0	1433.0	2803.0
NPDES	-114.7	-4.1	22.8	150.8
Highland Rim	-103.9	-184.9	-147.3	-219.0
Urban	306.1	30.2	-34.3	-257.0
Central Till Plain	957.3	1044.0	1048.0	1393.0
Forest to Urban	-228.9	28.6	91.6	197.1
Water	180.2	-56.0	-143.4	-344.5
Limestone	-128.2	-22.8	-4.8	-9.5
Network Density	152.2	149.7	217.3	245.4
Sum of Streams	-59.5	52.9	71.7	146.5
Grassland	-206.8	-177.4	-238.8	-397.5
Agriculture to Forest	-93.5	33.7	142.4	481.2
Temperature	647.3	322.3	210.0	-354.1
Shawnee Hills	-169.6	-240.8	-305.1	-700.8
CSOs	80.0	42.0	0.7	-88.2

Supplementary Table 7.12 (cont.) – Classification coefficients and constant for the quarter 2 geometric mean LDA classification equations

Variable	Cluster5	Cluster6	Cluster7
Constant	-567.3	-914.7	-611.1
Forest	1311.0	2307.0	1409.0
NPDES	81.4	71.0	13.0
Highland Rim	-126.5	-186.2	-220.5
Urban	-30.6	-133.8	-73.3
Central Till Plain	1003.0	1302.0	1130.0
Forest to Urban	53.9	128.9	171.3
Water	-102.5	-217.2	-185.4
Limestone	-6.8	-36.0	15.9
Network Density	192.4	207.7	202.3
Sum of Streams	53.3	97.0	93.6
Grassland	-219.0	-320.9	-209.7
Agriculture to Forest	153.1	258.0	107.0
Temperature	222.1	-36.0	215.2
Shawnee Hills	-270.3	-535.3	-275.8
CSOs	-0.4	-39.5	6.2

Supplementary Table 7.13 – Classification coefficients and constant for the quarter 3 mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-259.1	-254.1	-211.1	-269.1
Drainage Area	195.3	198.3	-47.5	192.1
Temperature	132.5	123.8	401.2	142.2
Urban	-105.7	-77.8	12.8	-117.6
Agriculture to Forest	124.6	142.8	14.5	125.2
Forest to Urban	92.9	64.6	8.3	94.8
Water	-74.3	-59.6	24.2	-78.7
Moderately Well Drained Soil	83.6	81.5	48.7	76.9
Central Till Plain	496.1	470.5	325.6	521.3
Limestone	42.6	39.1	10.8	40.3

Supplementary Table 7.13 (cont.) – Classification coefficients and constant for the quarter 3 mean LDA classification equations

Variable	Cluster5	Cluster6	Cluster7
Constant	-199.7	-319.0	-248.7
Drainage Area	66.2	310.4	183.2
Temperature	247.7	42.2	171.5
Urban	3.1	-97.8	-73.6
Agriculture to Forest	106.7	189.1	118.9
Forest to Urban	-11.0	77.9	61.7
Water	7.6	-78.4	-45.1
Moderately Well Drained Soil	51.5	117.6	94.3
Central Till Plain	355.4	479.9	435.5
Limestone	14.9	32.4	25.3

Supplementary Table 7.14 – Classification coefficients and constant for the quarter 3 median LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-147.8	-159.6	-149.8	-113.3	-157.3
Temperature	105.3	124.7	115.8	118.3	216.7
Longest Flow Path	-131.7	-170.3	-97.8	-136.7	-210.3
CSOs	88.3	98.1	102.4	52.5	91.2
NPDES	71.3	84.3	60.4	77.0	106.4
Agriculture to Forest	135.9	139.1	122.9	119.2	86.0
Sum of Streams	144.9	175.0	138.2	106.1	159.1
Cultivated Cropland	228.5	251.4	217.1	197.3	216.2
Poorly Drained Soil	-76.0	-94.4	-85.1	-54.7	-94.6
Wetlands	12.9	-7.1	6.2	21.1	-0.6

Supplementary Table 7.15 – Classification coefficients and constant for the quarter 3 trimmed mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-396.9	-471.0	-420.2	-377.4	-382.9
Temperature	438.6	565.7	430.4	598.7	446.4
Urban	-63.6	-2.3	-82.0	29.3	-35.1
Bluegrass	58.9	67.6	82.2	32.9	63.7
Central Till Plain	630.1	597.8	702.9	522.1	630.2
Agriculture to Forest	199.2	204.4	212.8	124.7	196.3
Forest to Urban	-173.8	-267.7	-154.6	-193.1	-181.1
Water	122.9	197.7	101.4	157.7	126.8
Network Density	255.4	264.1	275.8	216.5	235.6
Agriculture to Urban	185.3	256.8	186.1	179.0	173.0
Grassland	-128.1	-153.0	-145.1	-123.2	-124.4
Longest Flow Path	25.9	-94.1	-1.9	-96.8	-14.8
Poorly Drained Soil	37.5	72.9	20.2	59.6	49.4

Supplementary Table 7.16 – Classification coefficients and constant for the quarter 3 geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-2674.0	-3141.0	-2921.0	-3076.0	-2806.0
Interior Plateau	4774.0	5510.0	5227.0	5338.0	5150.0
Longest Flow Path	-214.6	-239.7	-230.9	-303.4	-262.6
Shawnee Hills	-413.2	-652.0	-638.8	-516.1	-567.4
Cultivated Cropland	1203.0	1228.0	1203.0	1296.0	1177.0
Highland Rim	-976.0	-1182.0	-1097.0	-1111.0	-1091.0
Moderately Well Drained	1057.0	1201.0	1161.0	1110.0	1093.0
Grassland	228.0	255.8	296.5	179.8	208.8
Temperature	-63.8	-343.6	85.3	-340.9	-289.2
Forest to Urban	657.7	697.3	691.2	699.1	617.7
Bluegrass	120.3	207.0	54.8	229.8	207.1
Network Density	77.6	123.0	27.4	159.1	157.8
CAFOs	364.9	458.4	355.3	434.3	417.2
Eastern Corn Belt	3138.0	3408.0	3262.0	3356.0	3222.0
Forest to Agriculture	676.7	801.2	680.5	813.2	756.7
Agriculture to Urban	665.9	794.4	652.4	804.0	783.3
Water	-94.4	-133.3	-81.5	-147.5	-107.7
Limestone	83.4	131.5	52.1	129.4	114.7

Supplementary Table 7.17 – Classification coefficients and constant for the quarter 4 mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Constant	-210.5	-163.0	-221.4	-246.3	-217.1
Shawnee Hills	-161.0	-73.3	-151.4	-147.7	-162.1
Interior Plateau	3.3	-47.9	2.2	-46.9	1.8
Central Till Plain	454.5	351.4	433.9	464.1	445.2
Urban	33.6	49.5	55.4	90.9	49.2
Forest to Urban	-74.5	-73.5	-97.7	-120.5	-86.5
Forest	545.5	471.4	566.7	633.2	554.5
Moderately Well Drained Soil	-57.6	-39.1	-60.3	-77.6	-54.2
Highland Rim	74.5	72.5	99.6	94.5	74.1
NPDES	2.3	-1.4	-9.9	-8.6	12.8

Supplementary Table 7.18 – Classification coefficients and constant for the quarter 4 median LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-136.8	-129.5	-324.9	-91.5
NPDES	183.1	230.2	436.2	197.0
Highland Rim	37.5	34.6	105.5	18.6
Interior Plateau	212.8	-6.4	-141.8	16.9
Shawnee Hills	-18.3	96.7	203.9	68.1
Cultivated Cropland	87.1	160.4	204.4	124.8
Wetlands	-32.8	-70.8	-91.8	-51.9
Agriculture to Urban	-70.4	-64.6	-161.8	-59.8
Impervious Surface	70.8	99.7	169.3	89.8
Drainage Area	10.3	66.3	122.6	57.1

Supplementary Table 7.18 (cont.) – Classification coefficients and constant for the quarter 4 median LDA classification equations

Variable	Cluster5	Cluster6	Cluster7
Constant	-156.5	-302.8	-123.7
NPDES	231.3	462.3	231.5
Highland Rim	92.3	35.3	28.6
Interior Plateau	-49.2	-10.7	2.1
Shawnee Hills	124.5	121.4	83.7
Cultivated Cropland	154.6	194.6	147.5
Wetlands	-54.7	-98.1	-50.3
Agriculture to Urban	-84.6	-166.3	-88.1
Impervious Surface	97.4	186.3	112.2
Drainage Area	79.1	88.9	66.0

Supplementary Table 7.19 – Classification coefficients and constant for the quarter 4 trimmed mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Constant	-106.8	-91.5	-109.6	-104.3	-55.2	-142.6
Interior Plateau	30.1	27.1	32.1	233.0	37.3	-35.9
Drainage Area	37.1	36.5	46.5	-20.1	25.1	77.6
Shawnee Hills	67.7	58.4	73.9	-38.1	38.0	115.5
Cultivated Cropland	203.0	173.1	199.8	119.1	126.3	204.8
NPDES	29.4	39.0	26.7	9.8	25.8	61.6
Urban to Forest	-36.4	-26.8	-41.3	-26.5	-17.4	-45.4
Forest to Urban	83.6	64.8	80.9	50.1	44.1	79.9
Water	-13.8	-1.0	-10.9	-1.7	6.8	-9.2

Supplementary Table 7.20 – Classification coefficients and constant for the quarter 4 geometric mean LDA classification equations

Variable	Cluster1	Cluster2	Cluster3	Cluster4
Constant	-63.1	-80.8	-78.7	-192.2
Forest	111.9	111.4	108.9	268.3
Drainage Area	41.3	47.0	55.2	-45.7
Interior Plateau	212.3	145.3	226.9	434.4
Shawnee Hills	-153.7	-105.6	-165.9	-291.6
Highland Rim	-69.0	0.2	-76.3	-69.7
NPDES	-12.9	-8.9	-5.6	-4.7
Moderately Well Drained	84.0	76.1	98.1	90.5
Urban	63.9	39.4	66.9	53.4

Supplementary Table 7.20 (cont.) – Classification coefficients and constant for the quarter 4 geometric mean LDA classification equations

Variable	Cluster5	Cluster6	Cluster7
Constant	-47.3	-130.6	-105.7
Forest	91.7	279.5	79.7
Drainage Area	34.0	-17.7	85.6
Interior Plateau	172.6	243.5	71.7
Shawnee Hills	-118.7	-224.0	-55.6
Highland Rim	-41.1	-78.6	4.9
NPDES	-3.8	16.2	18.1
Moderately Well Drained	66.9	83.7	81.6
Urban	44.0	49.5	38.2

LDA and SVM Cluster Prediction for the ECWMP Sites

Supplementary Table 8.1 – Cluster prediction and posterior probability error rate estimates for the Annual LDA models' classification of the ECWMP sites

ANNUAL	Geometric Mean (6)		Mean (5)		Median (5)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	2	0.386	5	0.985	3	1	4	0.618
ECWMP-02	5	1	3	0.999	5	0.62	2	0.887
ECWMP-03	5	0.957	2	0.793	1	1	4	0.562
ECWMP-04	5	0.995	3	1	2	1	2	0.991
ECWMP-05	5	1	1	0.999	1	1	4	0.65
ECWMP-06	5	0.969	5	0.999	3	0.67	4	0.327
ECWMP-07	5	0.973	5	1	3	1	1	0.431
ECWMP-08	2	0.695	1	0.975	1	1	1	0.379
ECWMP-09	5	0.998	5	1	3	1	1	0.493
ECWMP-10	5	0.998	5	1	3	1	1	0.386
ECWMP-11	5	0.995	5	1	3	1	1	0.438

Supplementary Table 8.2 – Cluster prediction and posterior probability error rate estimates for the Quarter 1 LDA models' classification of the ECWMP sites

QUARTER 1	Geometric Mean (5)		Mean (10)		Median (9)		Trimmed Mean (8)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	5	0.998	6	0.999	9	1	1	0.999
ECWMP-02	5	0.974	6	1	4	1	7	0.748
ECWMP-03	5	0.998	2	0.583	4	1	7	0.418
ECWMP-04	5	0.901	2	0.607	4	1	2	0.496
ECWMP-05	2	0.556	6	1	4	1	1	0.968
ECWMP-06	5	0.999	6	1	4	1	1	1
ECWMP-07	2	1	8	1	9	0.98	1	1
ECWMP-08	5	0.699	6	1	5	0.97	1	0.998
ECWMP-09	2	1	8	1	9	1	1	1
ECWMP-10	2	1	1	0.971	9	1	1	1
ECWMP-11	2	1	1	0.992	9	1	1	1

Supplementary Table 8.3 – Cluster prediction and posterior probability error rate estimates for the Quarter 2 LDA models' classification of the ECWMP sites

QUARTER 2	Geometric Mean (7)		Mean (5)		Median (5)		Trimmed Mean (7)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	2	1	3	0.721	2	1	4	1
ECWMP-02	2	1	4	0.999	5	0.99	7	0.999
ECWMP-03	2	1	3	0.535	3	0.41	1	1
ECWMP-04	2	1	3	0.978	5	1	3	0.997
ECWMP-05	2	1	3	1	3	1	7	1
ECWMP-06	2	0.959	3	0.996	2	1	7	1
ECWMP-07	3	0.588	3	1	2	1	4	1
ECWMP-08	2	1	3	1	2	1	3	1
ECWMP-09	7	0.6	3	1	2	1	4	1
ECWMP-10	7	0.548	3	1	2	1	4	1
ECWMP-11	2	0.999	3	1	2	1	4	1

Supplementary Table 8.4 – Cluster prediction and posterior probability error rate estimates for the Quarter 3 LDA models' classification of the ECWMP sites

QUARTER 3	Geometric Mean (5)		Mean (7)		Median (5)		Trimmed Mean (5)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	1	0.53	5	1	4	1	3	0.524
ECWMP-02	5	1	5	1	4	1	5	0.984
ECWMP-03	2	0.995	2	0.642	4	0.79	3	0.506
ECWMP-04	5	1	5	0.999	2	0.65	5	0.993
ECWMP-05	2	1	4	0.926	4	0.64	5	0.837
ECWMP-06	4	1	4	0.886	4	0.89	3	0.971
ECWMP-07	4	1	4	0.953	2	0.82	3	1
ECWMP-08	4	1	4	0.841	4	0.86	3	0.994
ECWMP-09	4	1	4	0.97	2	1	3	1
ECWMP-10	4	1	4	0.975	2	0.96	3	1
ECWMP-11	4	0.986	4	0.976	2	0.93	3	1

Supplementary Table 8.5 – Cluster prediction and posterior probability error rate estimates for the Quarter 4 LDA models' classification of the ECWMP sites

QUARTER 4	Geometric Mean (7)		Mean (5)		Median (7)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	5	0.994	1	0.977	4	0.96	1	0.959
ECWMP-02	5	0.618	4	0.957	4	1	5	0.97
ECWMP-03	5	0.833	1	0.654	2	0.47	2	0.697
ECWMP-04	5	0.924	4	0.961	4	0.59	5	0.828
ECWMP-05	5	0.999	1	0.813	2	1	2	0.819
ECWMP-06	5	0.997	1	0.986	4	0.76	1	0.871
ECWMP-07	5	0.997	1	0.989	2	0.94	1	0.99
ECWMP-08	5	0.995	1	0.983	4	0.47	1	0.973
ECWMP-09	5	0.999	1	0.994	2	1	1	0.993
ECWMP-10	5	0.999	1	0.994	2	1	1	0.994
ECWMP-11	5	1	1	0.996	2	0.99	1	0.993

Supplementary Table 8.6 – Cluster prediction and probability estimates for the Annual SVM models' classification of the ECWMP sites

ANNUAL	Geometric Mean(8)		Mean (3)		Median (7)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	5	0.431	1	0.882	2	0.436	4	0.451
ECWMP-02	5	0.383	1	0.918	4	0.392	6	0.402
ECWMP-03	4	0.496	1	0.934	4	0.535	6	0.556
ECWMP-04	4	0.473	1	0.929	4	0.506	6	0.529
ECWMP-05	5	0.438	1	0.907	2	0.445	4	0.456
ECWMP-06	5	0.565	1	0.928	2	0.572	4	0.589
ECWMP-07	5	0.659	1	0.871	2	0.666	4	0.692
ECWMP-08	4	0.437	1	0.921	4	0.464	6	0.48
ECWMP-09	5	0.673	1	0.85	2	0.679	4	0.706
ECWMP-10	5	0.687	1	0.836	2	0.693	4	0.718
ECWMP-11	5	0.656	1	0.844	2	0.662	4	0.688

Supplementary Table 8.7 – Cluster prediction and probability estimates for the Quarter 1 SVM models' classification of the ECWMP sites

QUARTER 1	Geometric Mean (9)		Mean (7)		Median (6)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	1	0.449	7	0.498	3	0.45	5	0.704
ECWMP-02	1	0.358	7	0.478	2	0.501	5	0.658
ECWMP-03	7	0.41	7	0.57	2	0.596	5	0.782
ECWMP-04	1	0.379	7	0.635	2	0.549	5	0.822
ECWMP-05	1	0.551	7	0.724	3	0.495	5	0.844
ECWMP-06	1	0.651	7	0.728	3	0.669	5	0.852
ECWMP-07	1	0.671	7	0.662	3	0.769	5	0.83
ECWMP-08	1	0.52	7	0.68	3	0.486	5	0.841
ECWMP-09	1	0.692	7	0.684	3	0.811	5	0.826
ECWMP-10	1	0.698	7	0.655	3	0.823	5	0.83
ECWMP-11	1	0.688	7	0.692	3	0.808	5	0.812

Supplementary Table 8.8 – Cluster prediction and probability estimates for the Quarter 2 SVM models' classification of the ECWMP sites

QUARTER 2	Geometric Mean (9)		Mean (9)		Median (8)		Trimmed Mean (6)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	3	0.422	6	0.31	4	0.633	2	0.545
ECWMP-02	3	0.375	6	0.298	4	0.683	2	0.499
ECWMP-03	3	0.474	6	0.414	4	0.699	2	0.801
ECWMP-04	3	0.491	6	0.427	4	0.642	2	0.781
ECWMP-05	3	0.479	6	0.508	4	0.753	2	0.721
ECWMP-06	3	0.465	6	0.448	4	0.801	2	0.742
ECWMP-07	3	0.463	5	0.326	4	0.523	2	0.628
ECWMP-08	3	0.498	6	0.422	4	0.566	2	0.746
ECWMP-09	3	0.464	5	0.336	4	0.499	2	0.591
ECWMP-10	3	0.458	5	0.326	4	0.463	2	0.568
ECWMP-11	3	0.467	5	0.336	4	0.533	2	0.575

Supplementary Table 8.9 – Cluster prediction and probability estimates for the Quarter 3 SVM models' classification of the ECWMP sites

QUARTER 3	Geometric Mean (7)		Mean (4)		Median (5)		Trimmed Mean (5)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	7	0.436	1	0.435	1	0.447	4	0.691
ECWMP-02	7	0.461	1	0.582	1	0.561	4	0.784
ECWMP-03	7	0.427	3	0.552	3	0.491	4	0.841
ECWMP-04	7	0.446	3	0.48	3	0.517	4	0.847
ECWMP-05	7	0.491	1	0.526	1	0.529	4	0.872
ECWMP-06	7	0.501	1	0.545	1	0.526	4	0.852
ECWMP-07	7	0.512	1	0.636	1	0.668	4	0.807
ECWMP-08	7	0.485	1	0.487	3	0.488	4	0.858
ECWMP-09	7	0.506	1	0.63	1	0.693	4	0.789
ECWMP-10	7	0.517	1	0.665	1	0.771	4	0.807
ECWMP-11	7	0.491	1	0.596	1	0.651	4	0.76

Supplementary Table 8.10 – Cluster prediction and probability estimates for the Quarter 4 SVM models' classification of the ECWMP sites

QUARTER 4	Geometric Mean (5)		Mean (5)		Median (6)		Trimmed Mean (7)	
Station name	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**	**Cluster**	**Est.**
ECWMP-01	5	0.346	3	0.404	6	0.534	2	0.448
ECWMP-02	5	0.345	3	0.483	6	0.578	2	0.522
ECWMP-03	1	0.655	2	0.576	5	0.524	1	0.469
ECWMP-04	1	0.591	2	0.494	5	0.47	1	0.441
ECWMP-05	1	0.443	3	0.477	6	0.589	2	0.449
ECWMP-06	5	0.457	3	0.483	6	0.603	2	0.479
ECWMP-07	5	0.471	3	0.546	6	0.715	2	0.576
ECWMP-08	1	0.441	3	0.456	6	0.562	2	0.411
ECWMP-09	5	0.439	3	0.532	6	0.718	2	0.566
ECWMP-10	5	0.429	3	0.539	6	0.739	2	0.57
ECWMP-11	5	0.424	3	0.52	6	0.702	2	0.549

ECWMP Cluster Range Accuracy

Supplementary Table 9.1 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the annual datasets; highlighted values classified the highest percentage of variables within the specified range among different models for a given station

	Annual SVM Model Cluster Range Accuracy				**Annual LDA Model Cluster Range Accuracy**			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	78.6	71.4	85.7	85.7	64.3	57.1	85.7	78.6
ECWMP-02	71.4	92.9	78.6	78.6	71.4	85.7	71.4	35.7
ECWMP-03	78.6	71.4	78.6	64.3	64.3	78.6	78.6	85.7
ECWMP-04	78.6	85.7	85.7	78.6	78.6	71.4	57.1	28.6
ECWMP-05	64.3	71.4	78.6	57.1	71.4	14.3	85.7	71.4
ECWMP-06	78.6	71.4	78.6	78.6	78.6	57.1	78.6	71.4
ECWMP-07	78.6	71.4	71.4	85.7	78.6	35.7	78.6	71.4
ECWMP-08	85.7	71.4	71.4	78.6	85.7	50.0	71.4	78.6
ECWMP-09	85.7	57.1	71.4	85.7	85.7	35.7	85.7	71.4
ECWMP-10	78.6	57.1	85.7	71.4	78.6	50.0	85.7	64.3
ECWMP-11	64.3	57.1	64.3	64.3	57.1	50.0	71.4	57.1

Supplementary Table 9.2 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the quarter 1 datasets; highlighted values classified the highest percentage of variables within the specified range among different models for a given station

	Quarter 1 SVM Cluster Range Accuracy				Quarter 1 LDA Cluster Range Accuracy			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	57.1	57.1	85.7	64.3	50.0	0.0	14.3	50.0
ECWMP-02	57.1	64.3	42.9	78.6	71.4	42.9	57.1	78.6
ECWMP-03	50.0	78.6	42.9	64.3	85.7	78.6	64.3	71.4
ECWMP-04	71.4	64.3	35.7	78.6	64.3	64.3	78.6	28.6
ECWMP-05	57.1	71.4	57.1	50.0	64.3	35.7	71.4	28.6
ECWMP-06	57.1	50.0	64.3	71.4	64.3	21.4	57.1	57.1
ECWMP-07	71.4	57.1	71.4	64.3	78.6	14.3	14.3	50.0
ECWMP-08	71.4	57.1	78.6	71.4	50.0	28.6	78.6	64.3
ECWMP-09	64.3	57.1	71.4	71.4	78.6	7.1	14.3	50.0
ECWMP-10	71.4	64.3	64.3	71.4	71.4	7.1	14.3	57.1
ECWMP-11	71.4	64.3	57.1	71.4	78.6	14.3	21.4	57.1

Supplementary Table 9.3 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the quarter 2 datasets; highlighted values classified the highest percentage of variables within the specified range among different models for a given station

	Quarter 2 SVM Cluster Range Accuracy				Quarter 2 LDA Cluster Range Accuracy			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	78.6	35.7	50.0	71.4	42.9	71.4	71.4	28.6
ECWMP-02	50.0	21.4	28.6	64.3	42.9	92.9	71.4	35.7
ECWMP-03	50.0	42.9	50.0	64.3	71.4	57.1	71.4	78.6
ECWMP-04	71.4	35.7	50.0	78.6	57.1	71.4	78.6	42.9
ECWMP-05	35.7	35.7	50.0	50.0	71.4	28.6	71.4	21.4
ECWMP-06	57.1	35.7	28.6	78.6	42.9	42.9	42.9	50.0
ECWMP-07	50.0	21.4	28.6	57.1	35.7	42.9	50.0	28.6
ECWMP-08	42.9	28.6	28.6	64.3	35.7	57.1	64.3	35.7
ECWMP-09	50.0	21.4	21.4	64.3	28.6	35.7	28.6	28.6
ECWMP-10	64.3	35.7	42.9	64.3	64.3	50.0	57.1	42.9
ECWMP-11	50.0	21.4	42.9	57.1	42.9	42.9	50.0	28.6

Supplementary Table 9.4 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the quarter 3 datasets; highlighted values classified the highest percentage of variables within the specified range among different models for a given station

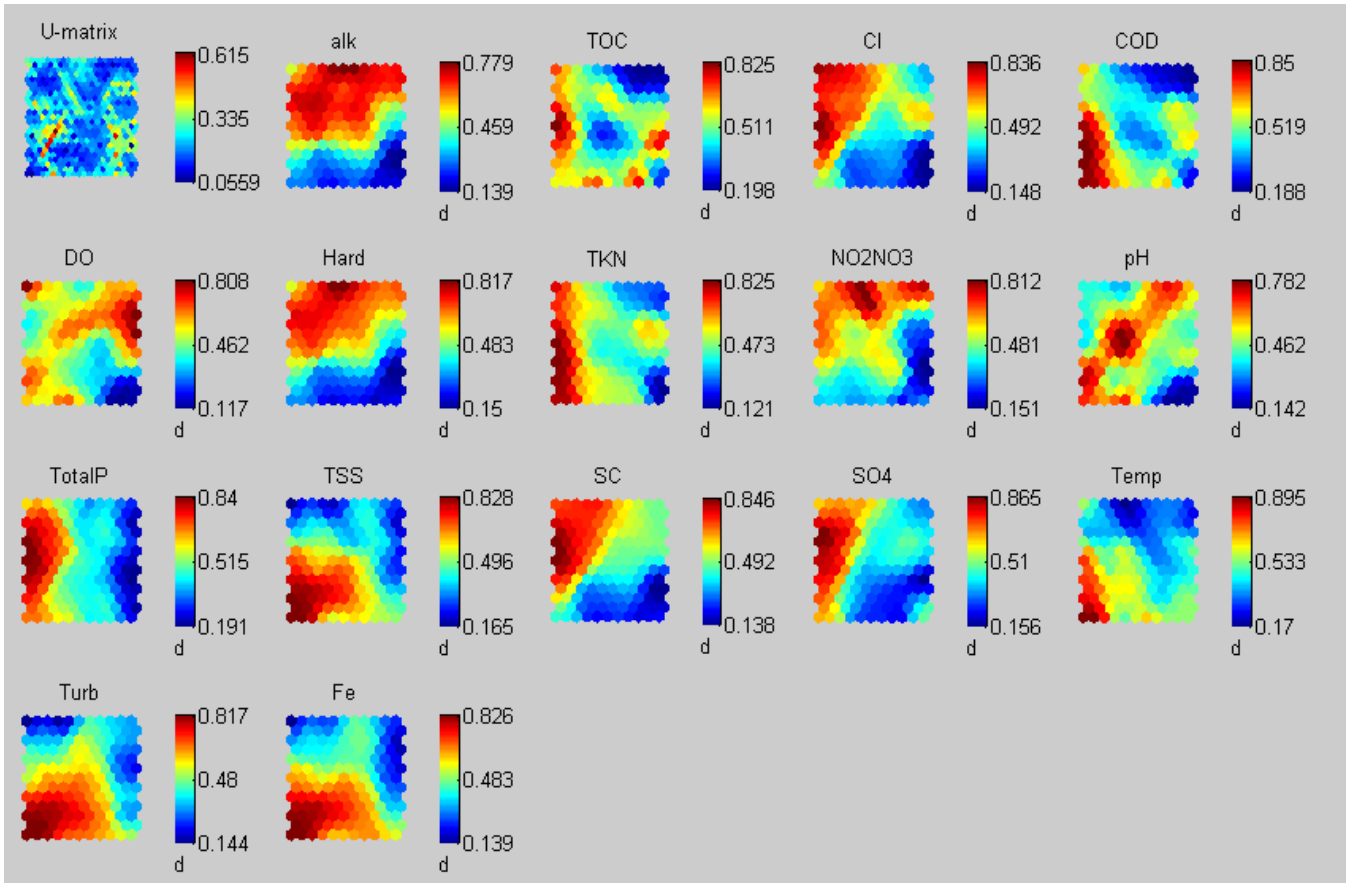
	Quarter 3 SVM Cluster Range Accuracy				**Quarter 3 LDA Cluster Range Accuracy**			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	92.9	85.7	42.9	85.7	42.9	50.0	71.4	85.7
ECWMP-02	50.0	42.9	35.7	57.1	42.9	57.1	92.9	42.9
ECWMP-03	42.9	64.3	28.6	57.1	78.6	57.1	35.7	35.7
ECWMP-04	64.3	78.6	35.7	64.3	71.4	50.0	28.6	71.4
ECWMP-05	50.0	57.1	64.3	57.1	71.4	21.4	28.6	57.1
ECWMP-06	57.1	64.3	64.3	71.4	57.1	21.4	35.7	57.1
ECWMP-07	42.9	50.0	28.6	50.0	50.0	28.6	28.6	42.9
ECWMP-08	64.3	78.6	50.0	78.6	64.3	50.0	78.6	71.4
ECWMP-09	50.0	42.9	28.6	35.7	71.4	21.4	42.9	35.7
ECWMP-10	50.0	42.9	21.4	42.9	64.3	28.6	50.0	35.7
ECWMP-11	42.9	50.0	28.6	42.9	42.9	35.7	35.7	42.9

Supplementary Table 9.5 – The percentage of water quality variables from the ECWMP dataset that fell within the range of the cluster to which it was assigned for the quarter 4 datasets; highlighted values classified the highest percentage of variables within the specified range among different models for a given station

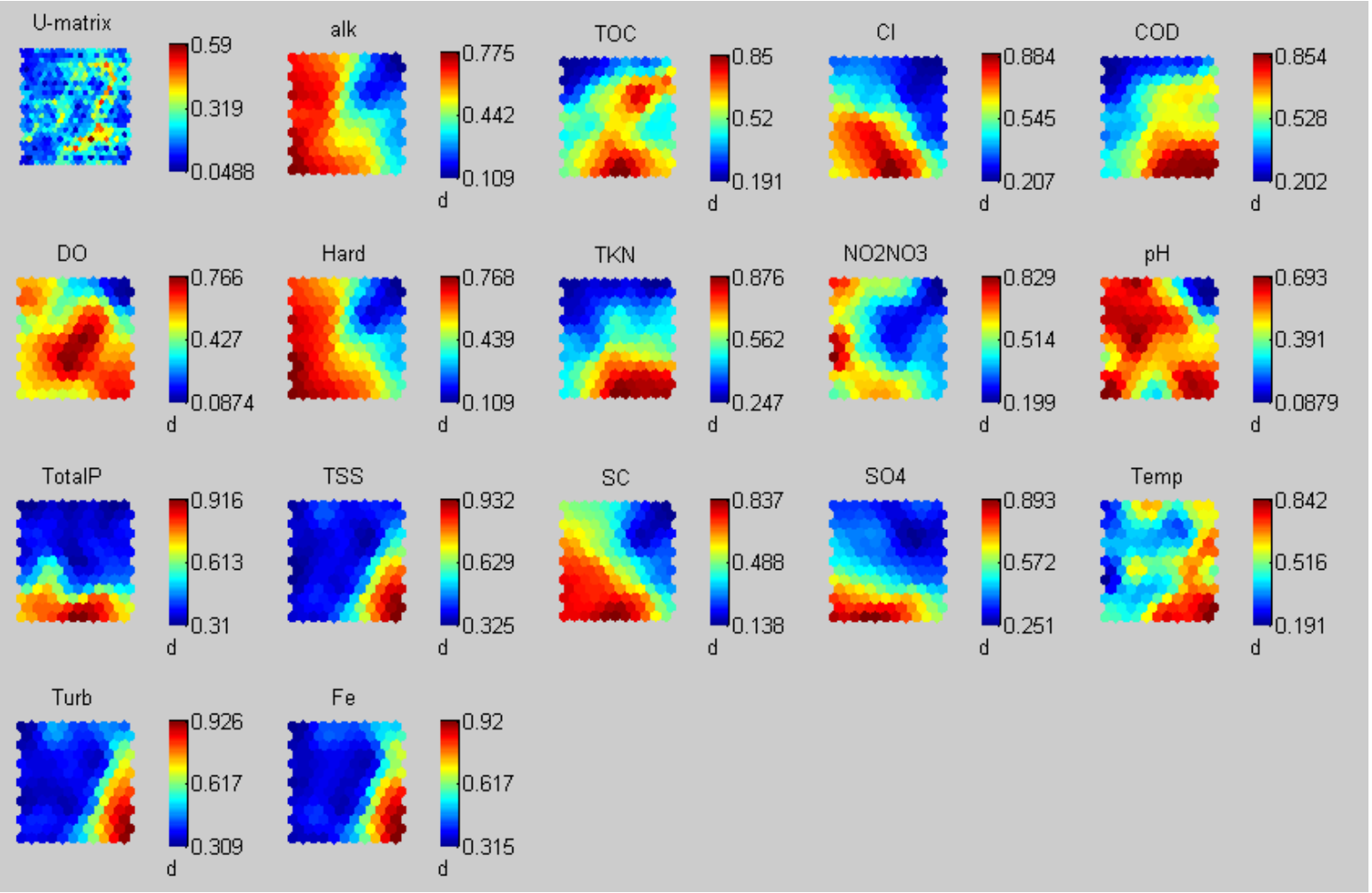
	Quarter 4 SVM Cluster Range Accuracy				Quarter 4 LDA Cluster Range Accuracy			
Station	Geometric Mean SVM	Mean SVM	Median SVM	Trimmed Mean SVM	Geometric Mean LDA	Mean LDA	Median LDA	Trimmed Mean LDA
ECWMP-01	42.9	50.0	42.9	50.0	35.7	50.0	50.0	42.9
ECWMP-02	57.1	42.9	50.0	50.0	57.1	78.6	64.3	64.3
ECWMP-03	35.7	64.3	57.1	57.1	35.7	42.9	35.7	42.9
ECWMP-04	50.0	64.3	50.0	50.0	42.9	57.1	78.6	50.0
ECWMP-05	50.0	35.7	42.9	35.7	28.6	57.1	35.7	50.0
ECWMP-06	42.9	57.1	50.0	50.0	35.7	50.0	57.1	35.7
ECWMP-07	50.0	64.3	57.1	42.9	42.9	42.9	42.9	28.6
ECWMP-08	7.1	71.4	64.3	64.3	42.9	42.9	71.4	35.7
ECWMP-09	28.6	35.7	35.7	42.9	35.7	35.7	28.6	28.6
ECWMP-10	50.0	50.0	42.9	42.9	28.6	28.6	35.7	35.7
ECWMP-11	35.7	35.7	35.7	35.7	35.7	35.7	14.3	35.7

APPENDIX B – SUPPLEMENTARY FIGURES

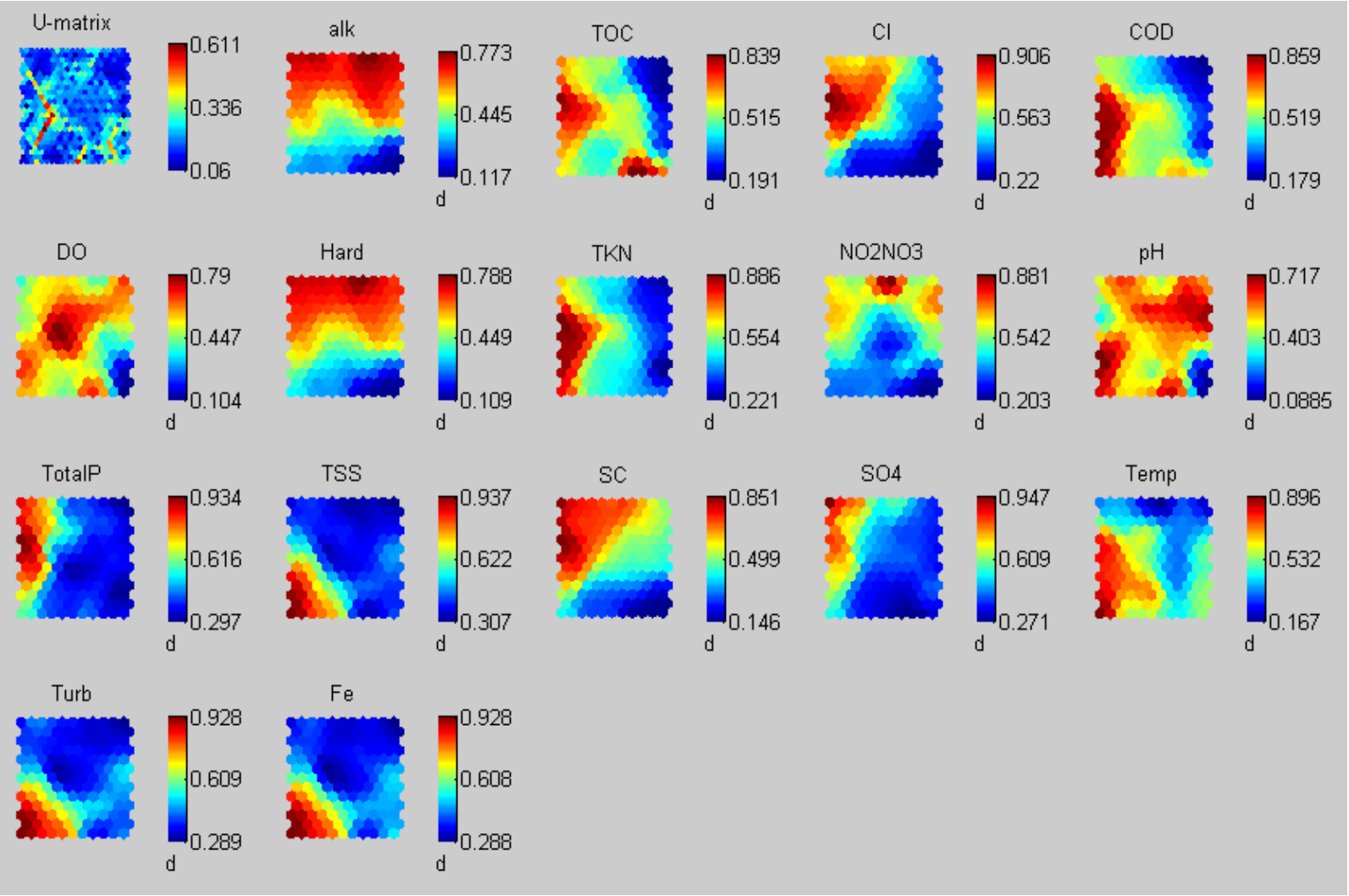
SOM Variable Component Maps



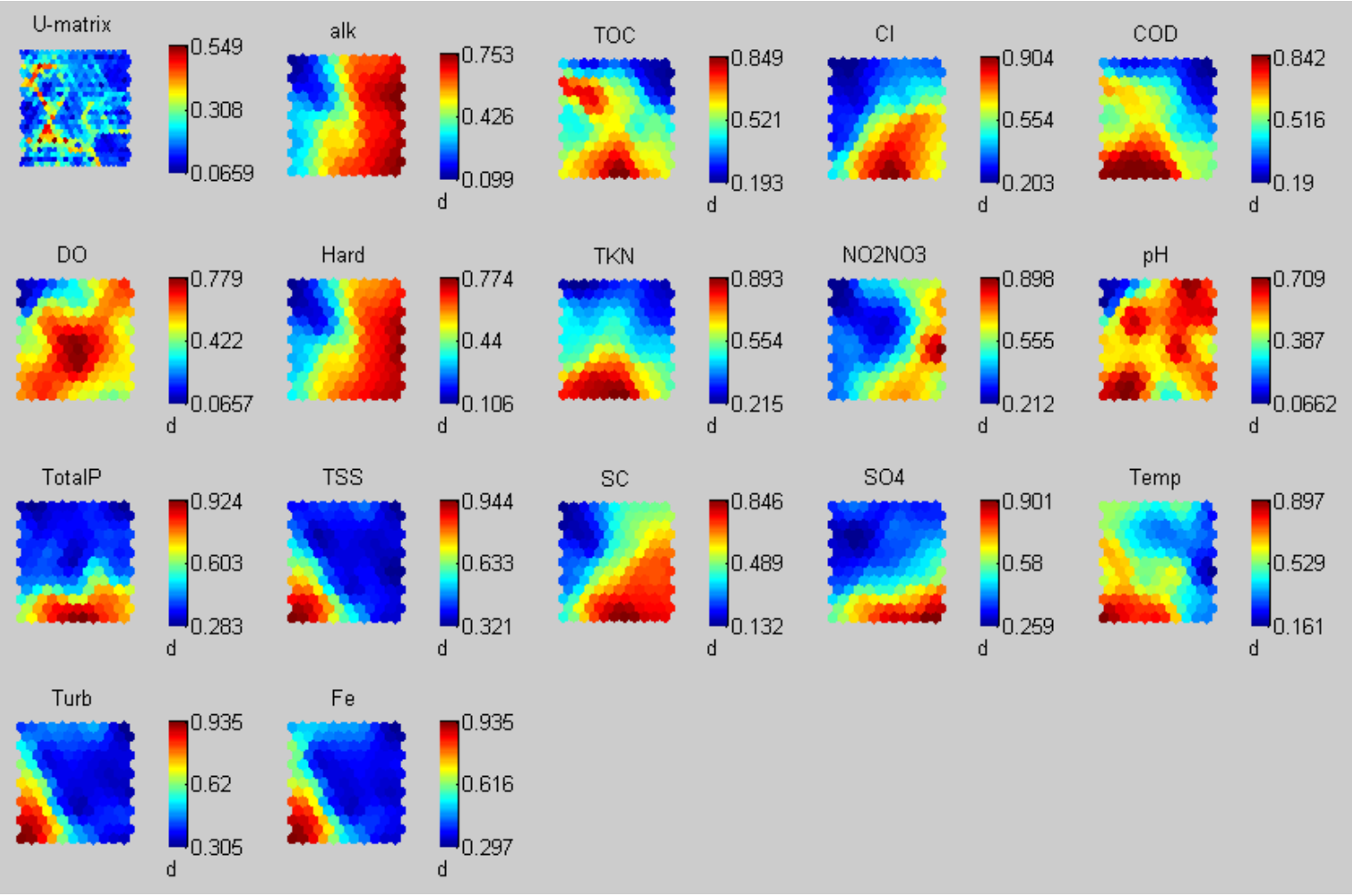
Supplementary Figure 1.1 – Annual Mean Dataset Component Maps



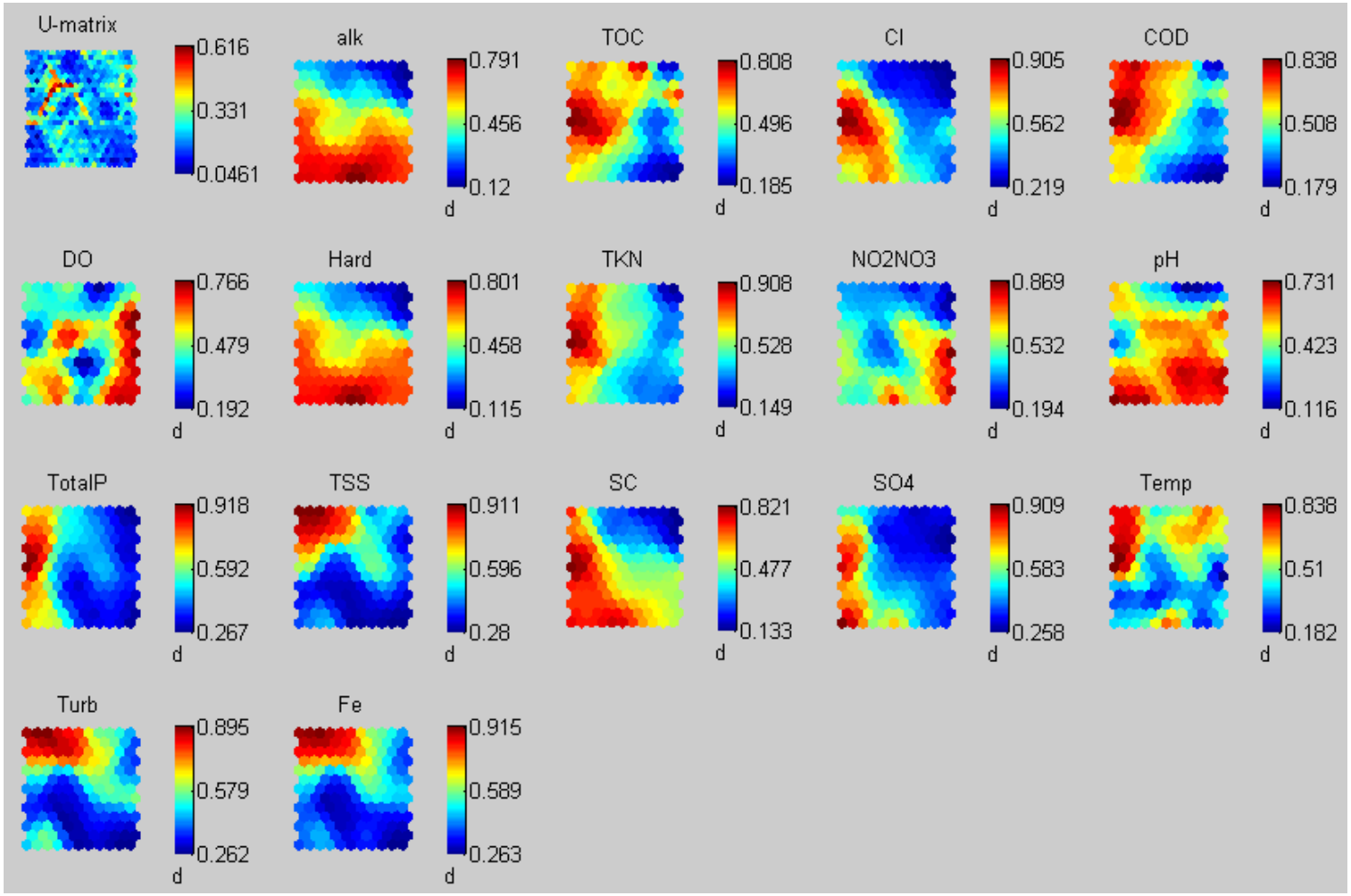
Supplementary Figure 1.2 – Annual Median Dataset Component Maps



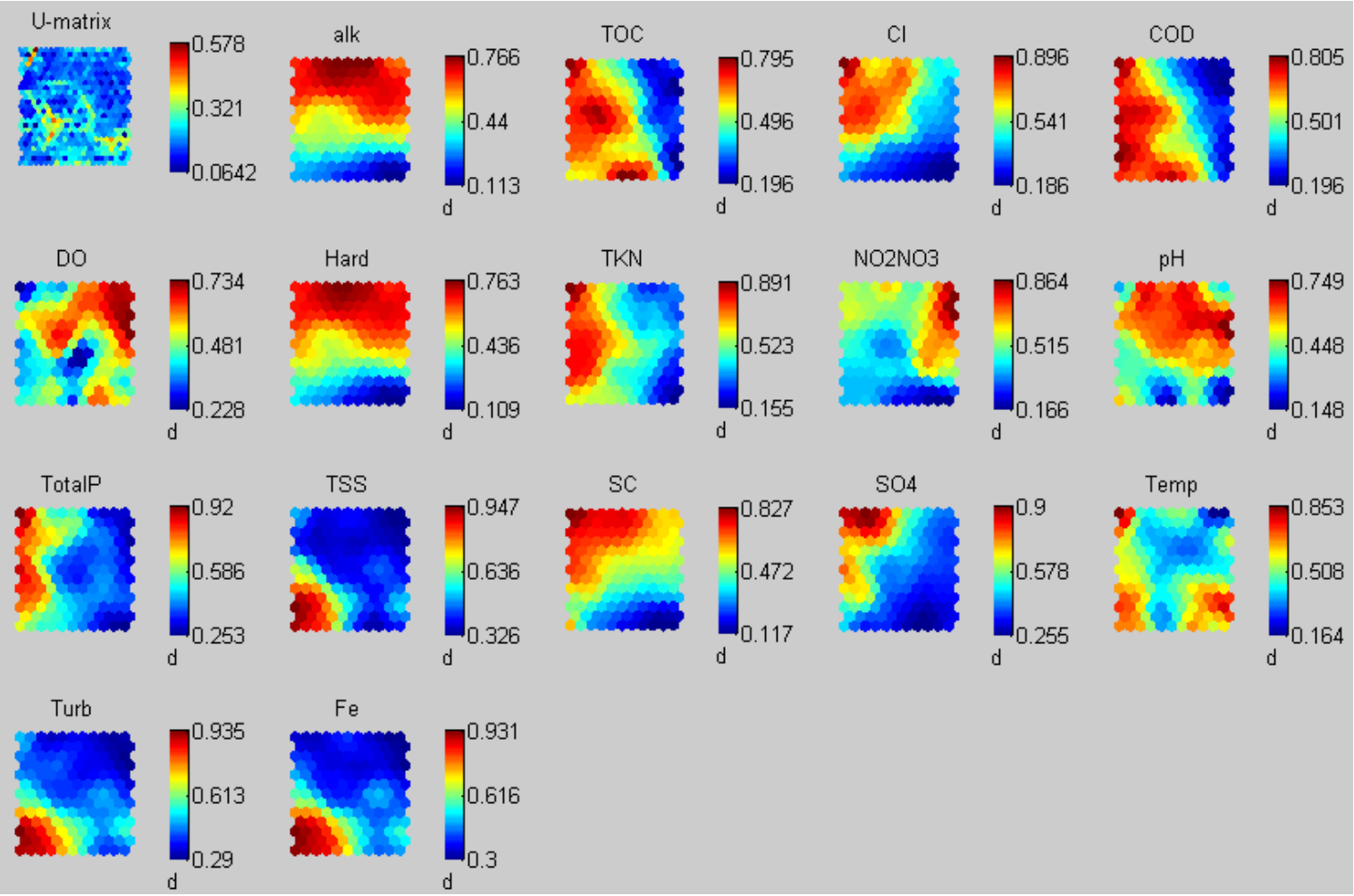
Supplementary Figure 1.3 – Annual Trimmed Mean Dataset Component Maps



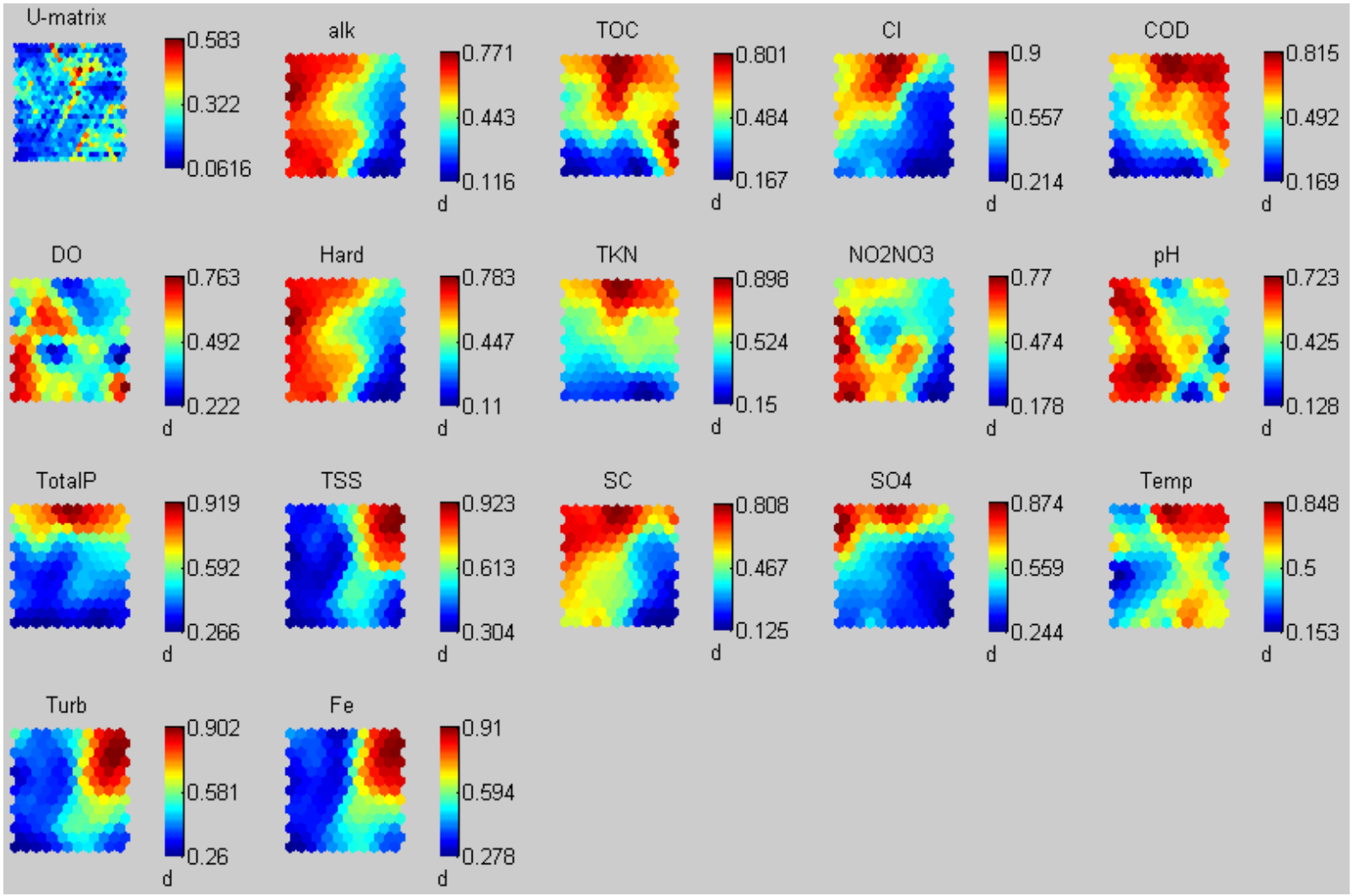
Supplementary Figure 1.4 – Annual Geometric Mean Dataset Component Maps



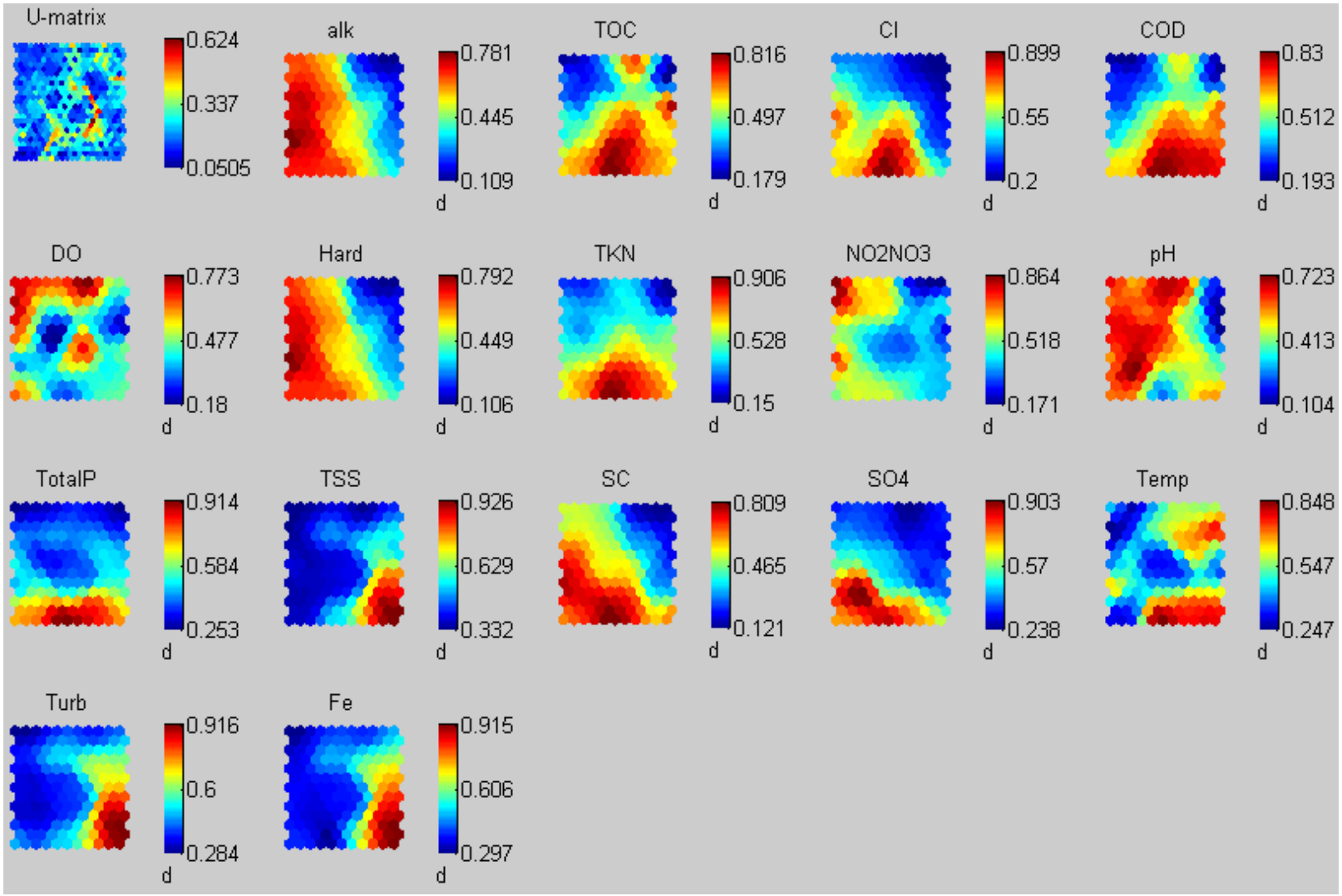
Supplementary Figure 1.5 – Quarter 1 Mean Dataset Component Maps



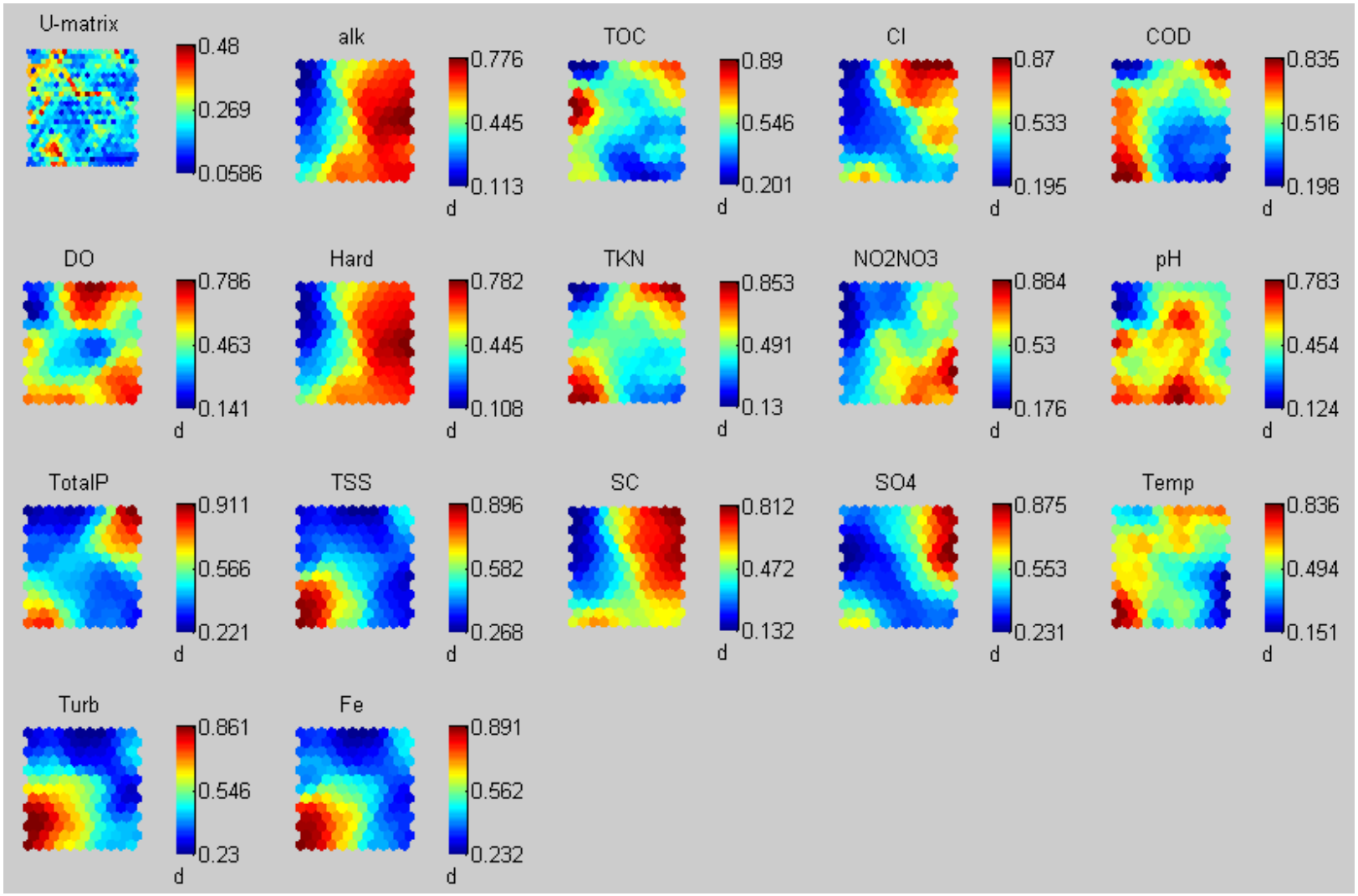
Supplementary Figure 1.6 – Quarter 1 Median Dataset Component Maps



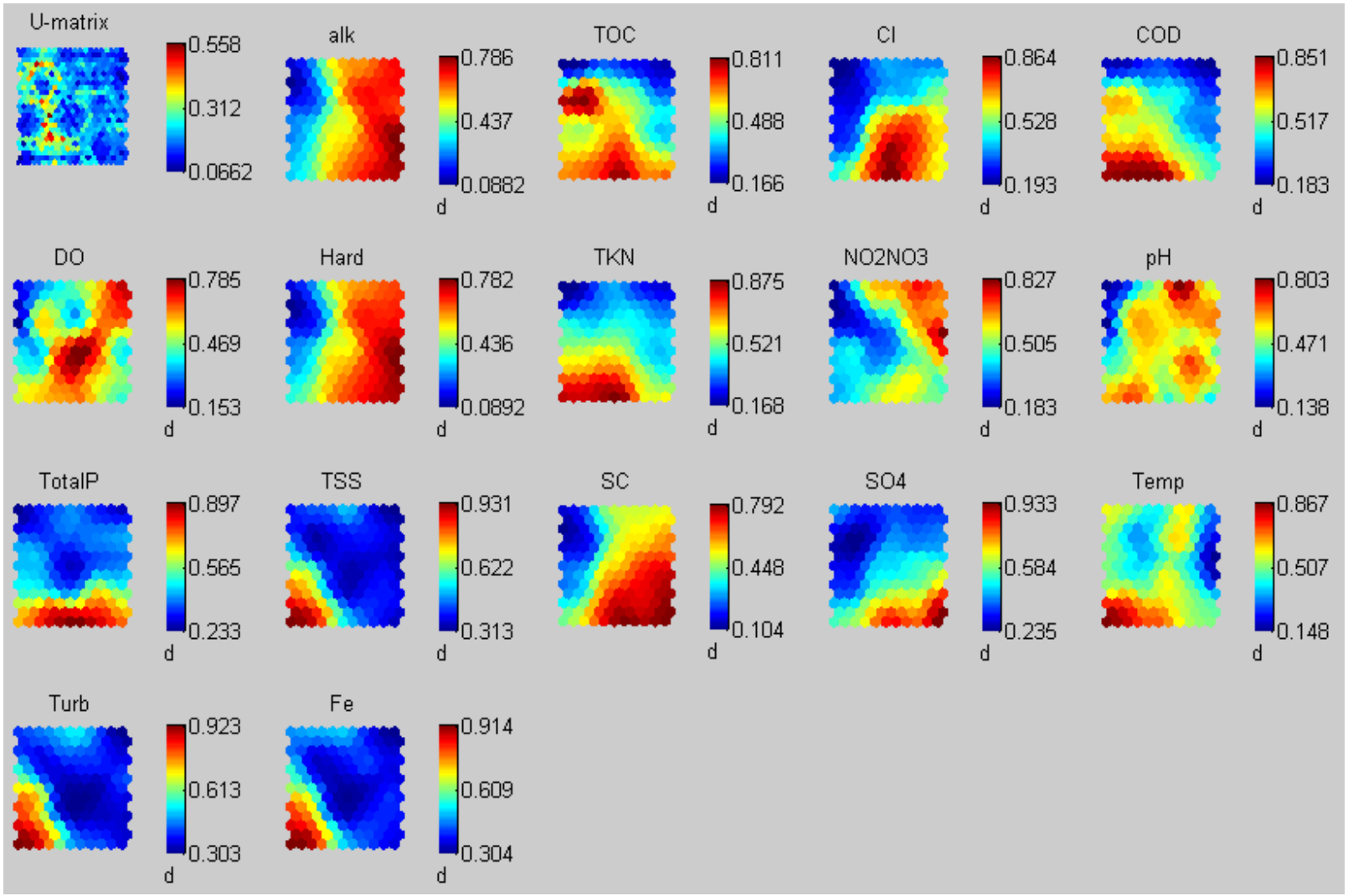
Supplementary Figure 1.7 – Quarter 1 Trimmed Mean Dataset Component Maps



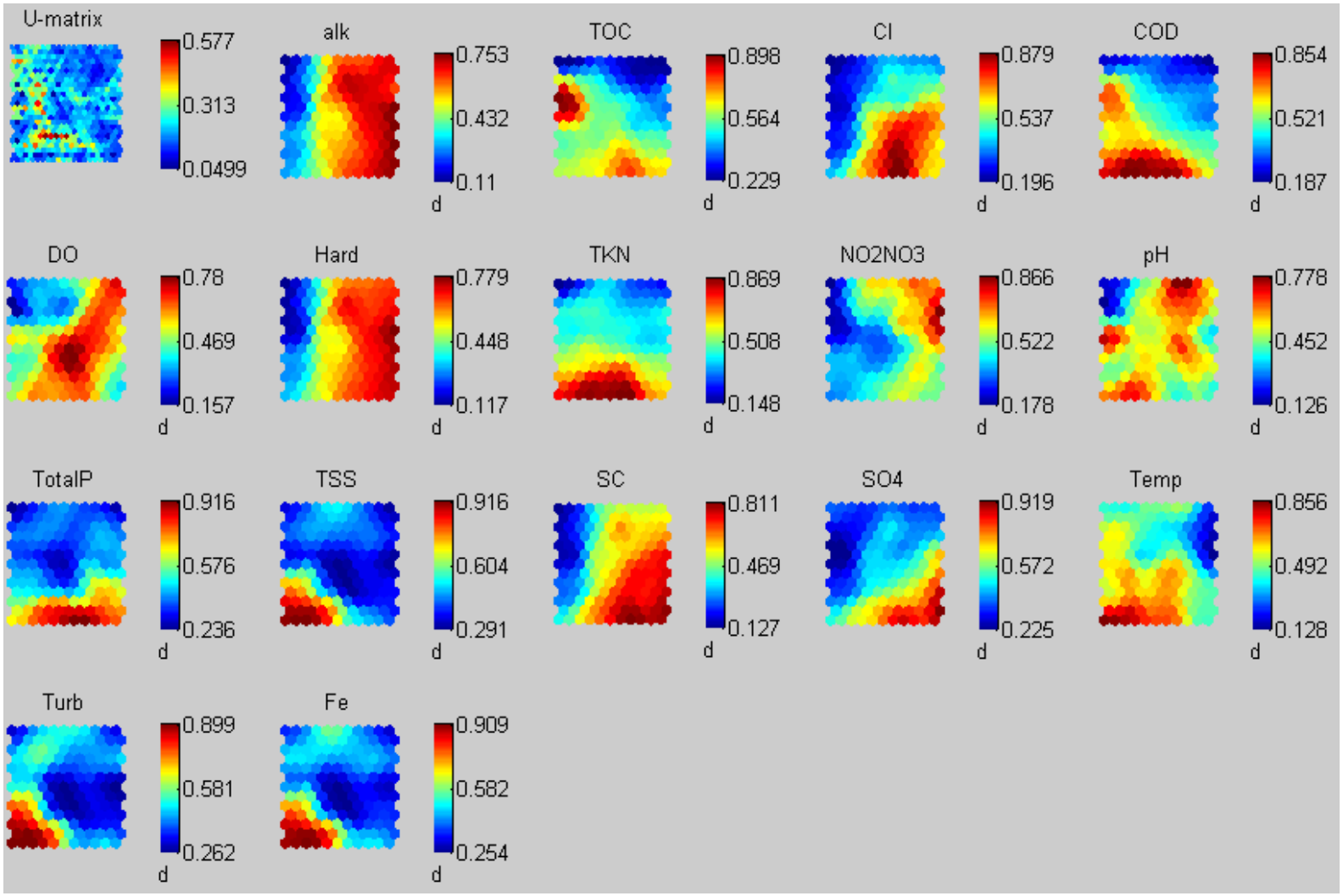
Supplementary Figure 1.8 – Quarter 1 Geometric Mean Dataset Component Maps



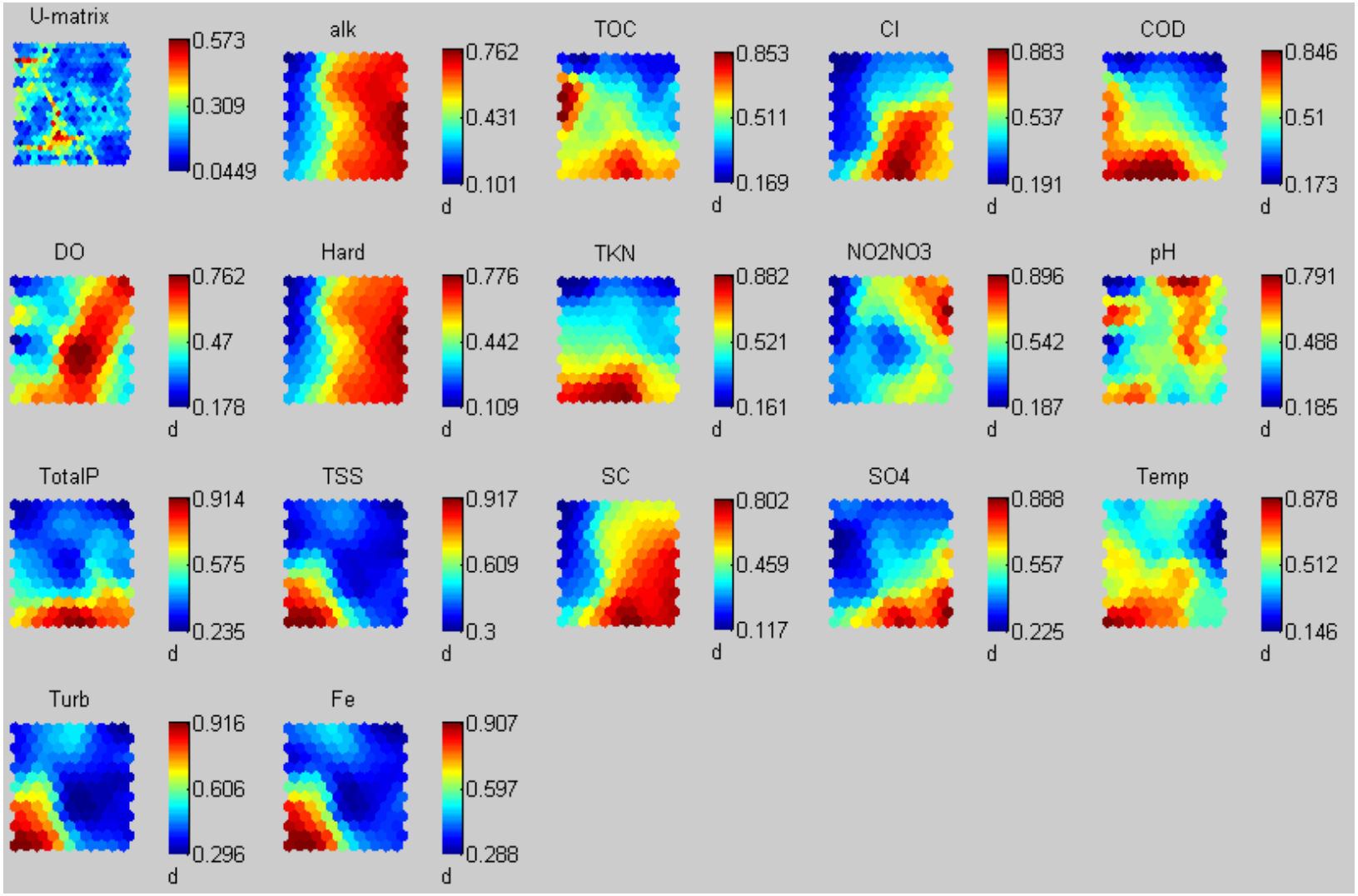
Supplementary Figure 1.9 – Quarter 2 Mean Dataset Component Maps



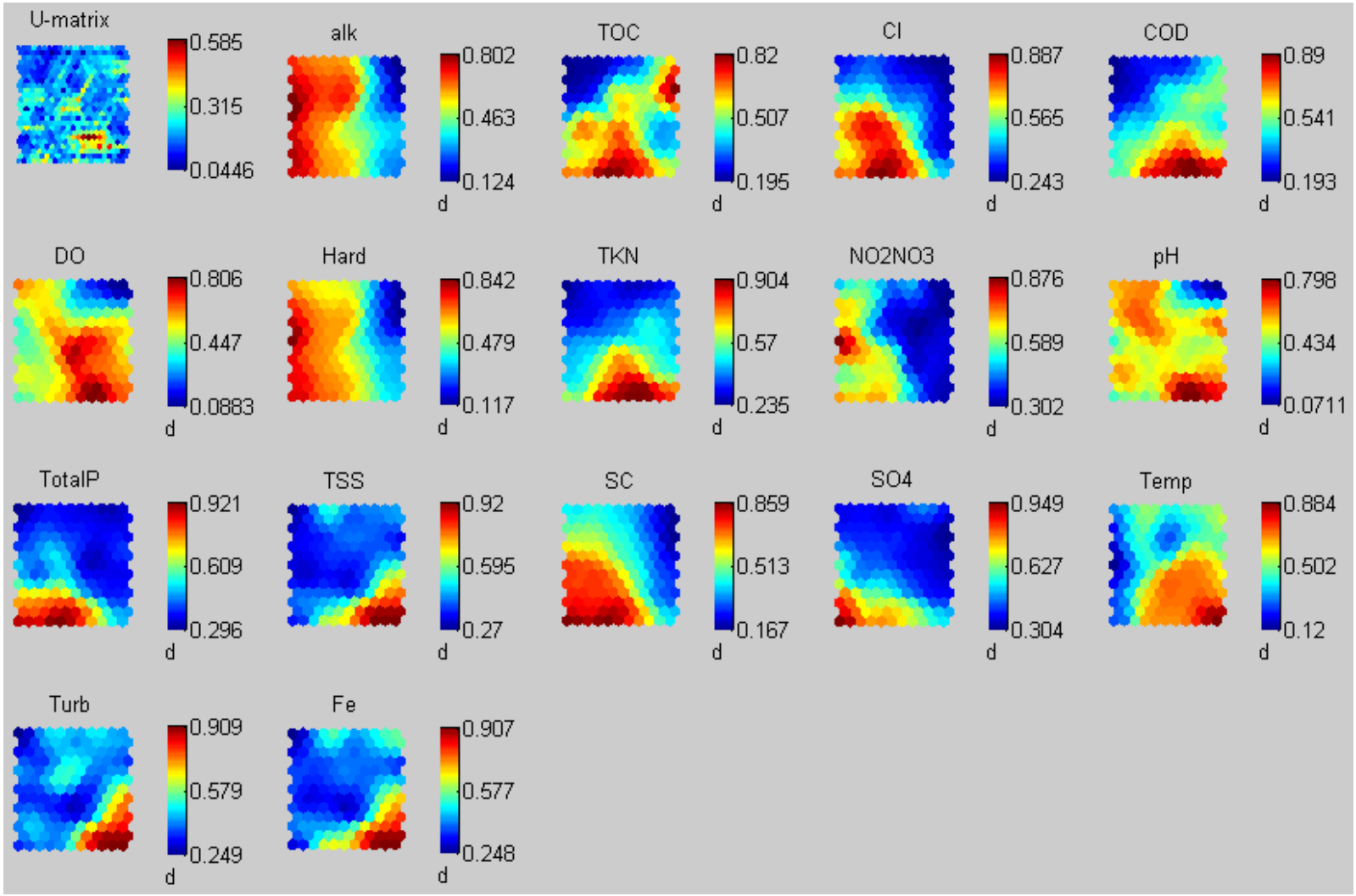
Supplementary Figure 1.10 – Quarter 2 Median Dataset Component Maps



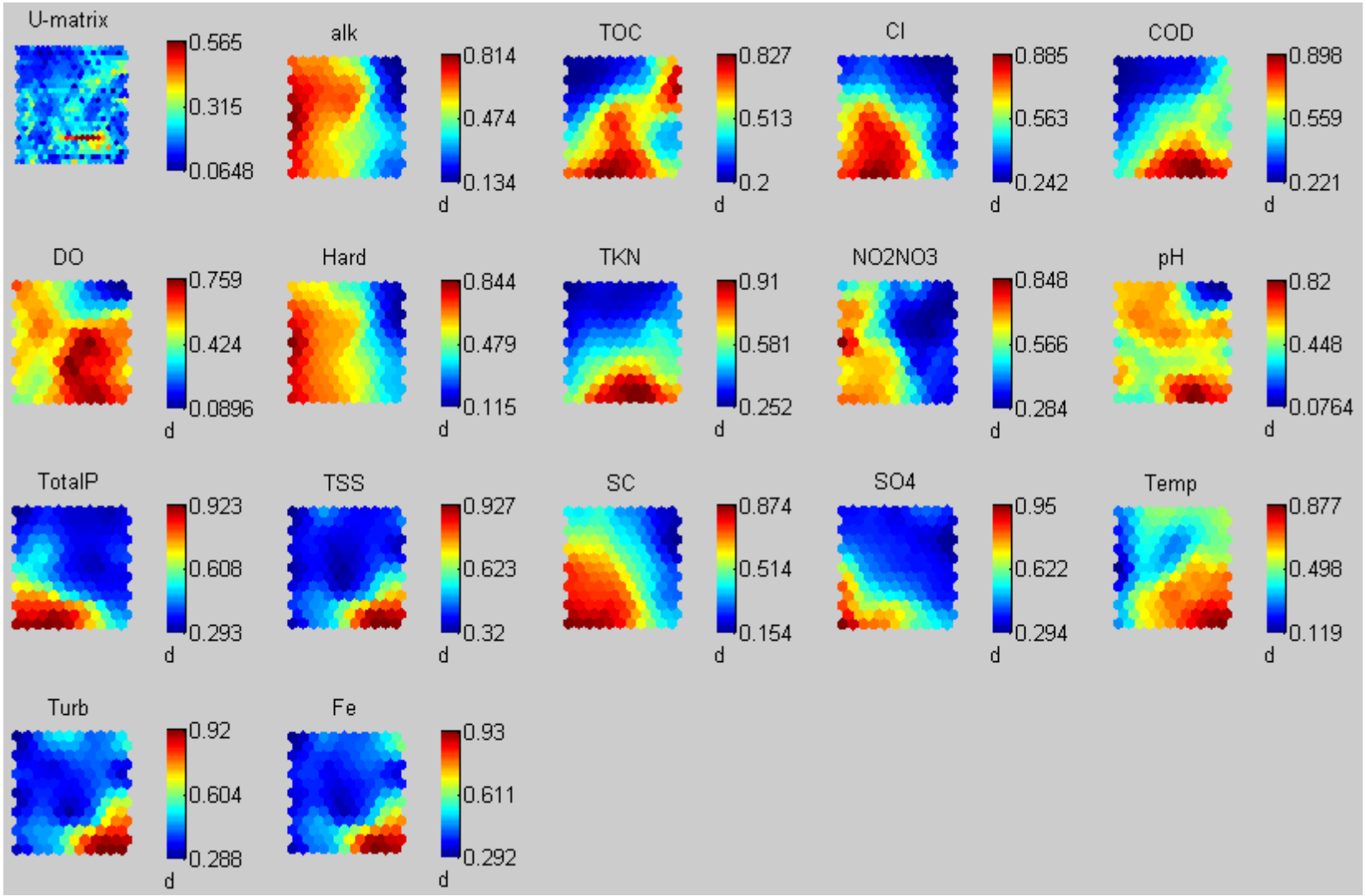
Supplementary Figure 1.11 – Quarter 2 Trimmed Mean Dataset Component Maps



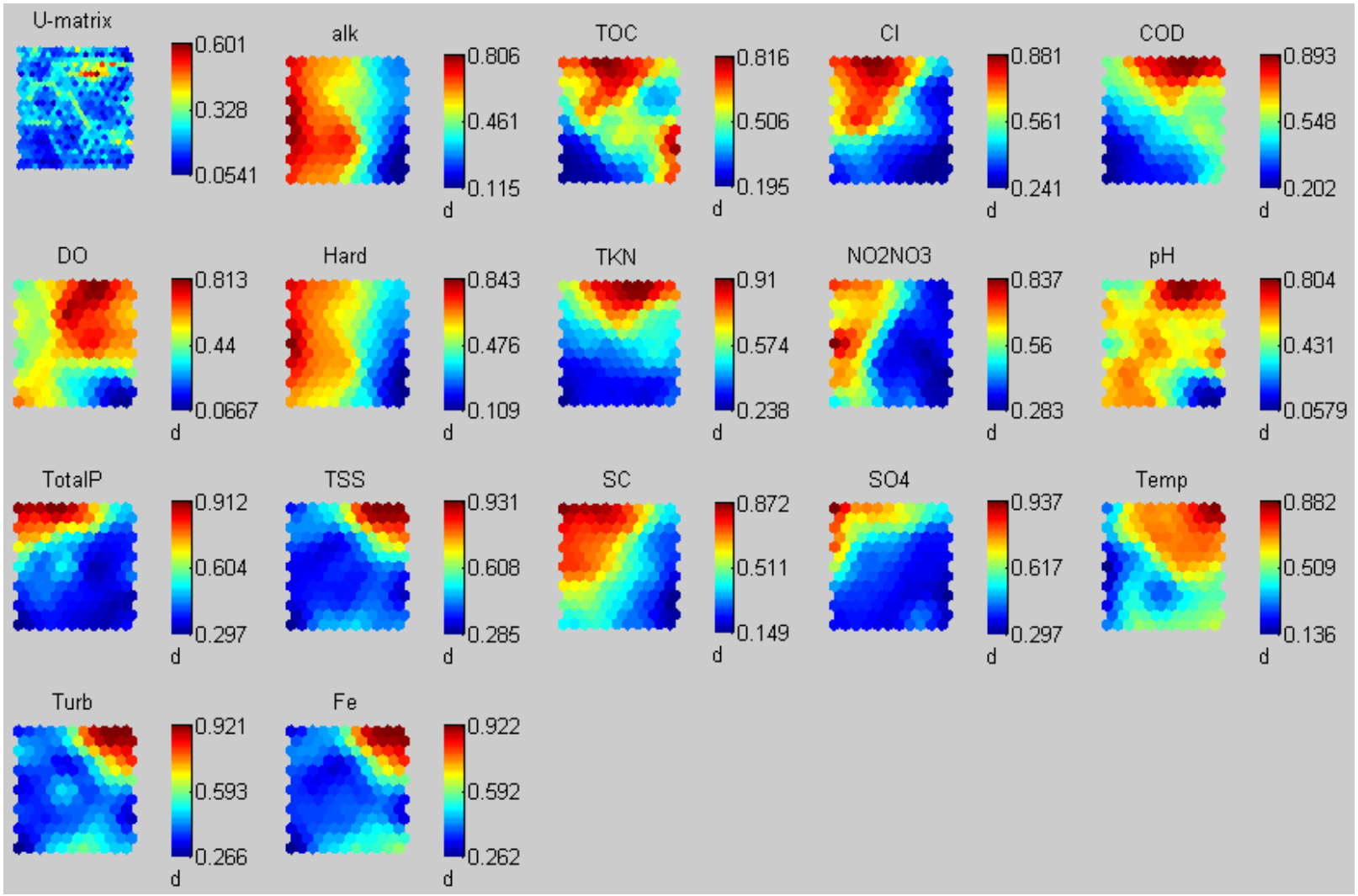
Supplementary Figure 1.12 – Quarter 2 Geometric Mean Dataset Component Maps



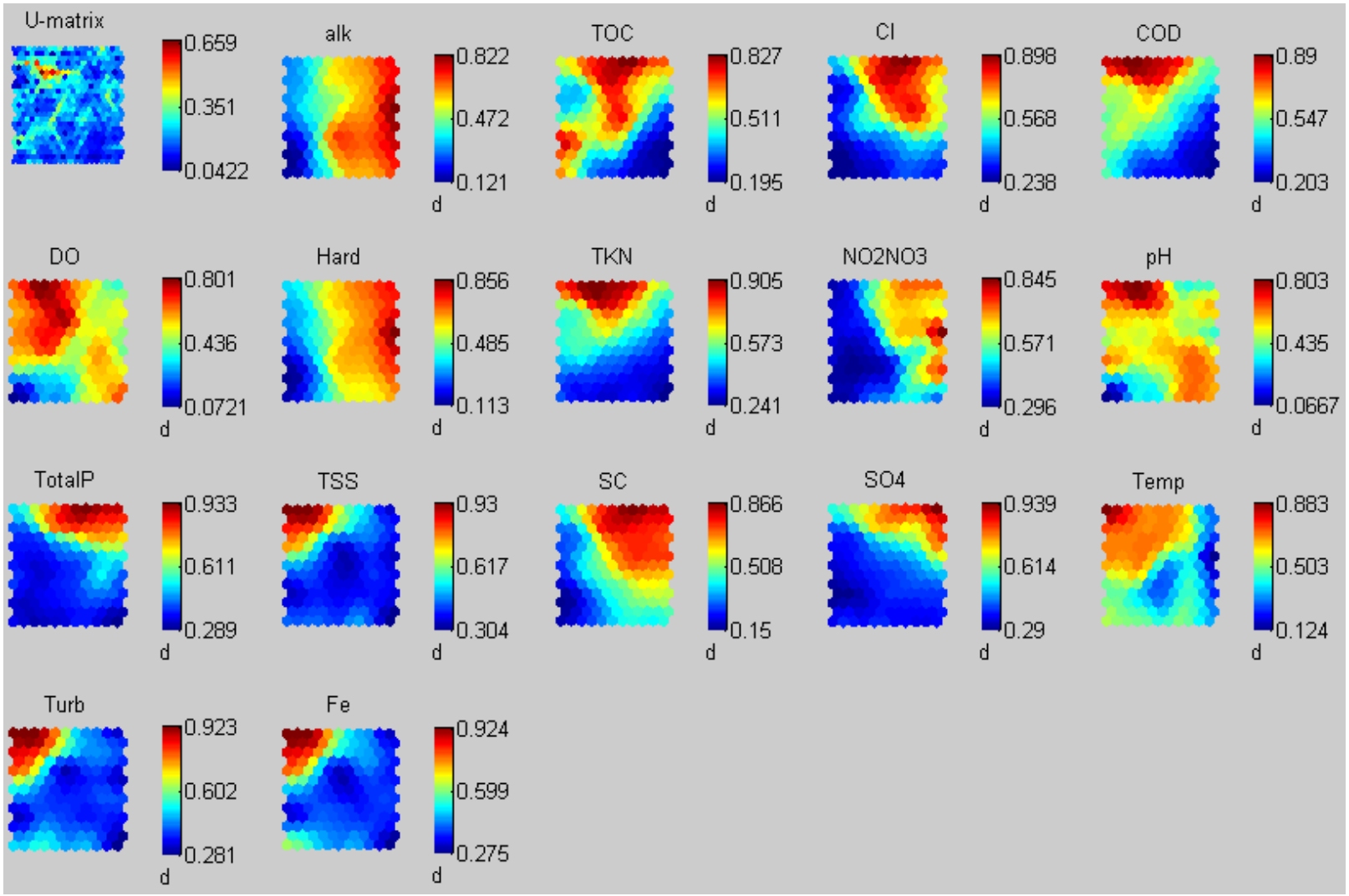
Supplementary Figure 1.13 – Quarter 3 Mean Dataset Component Maps



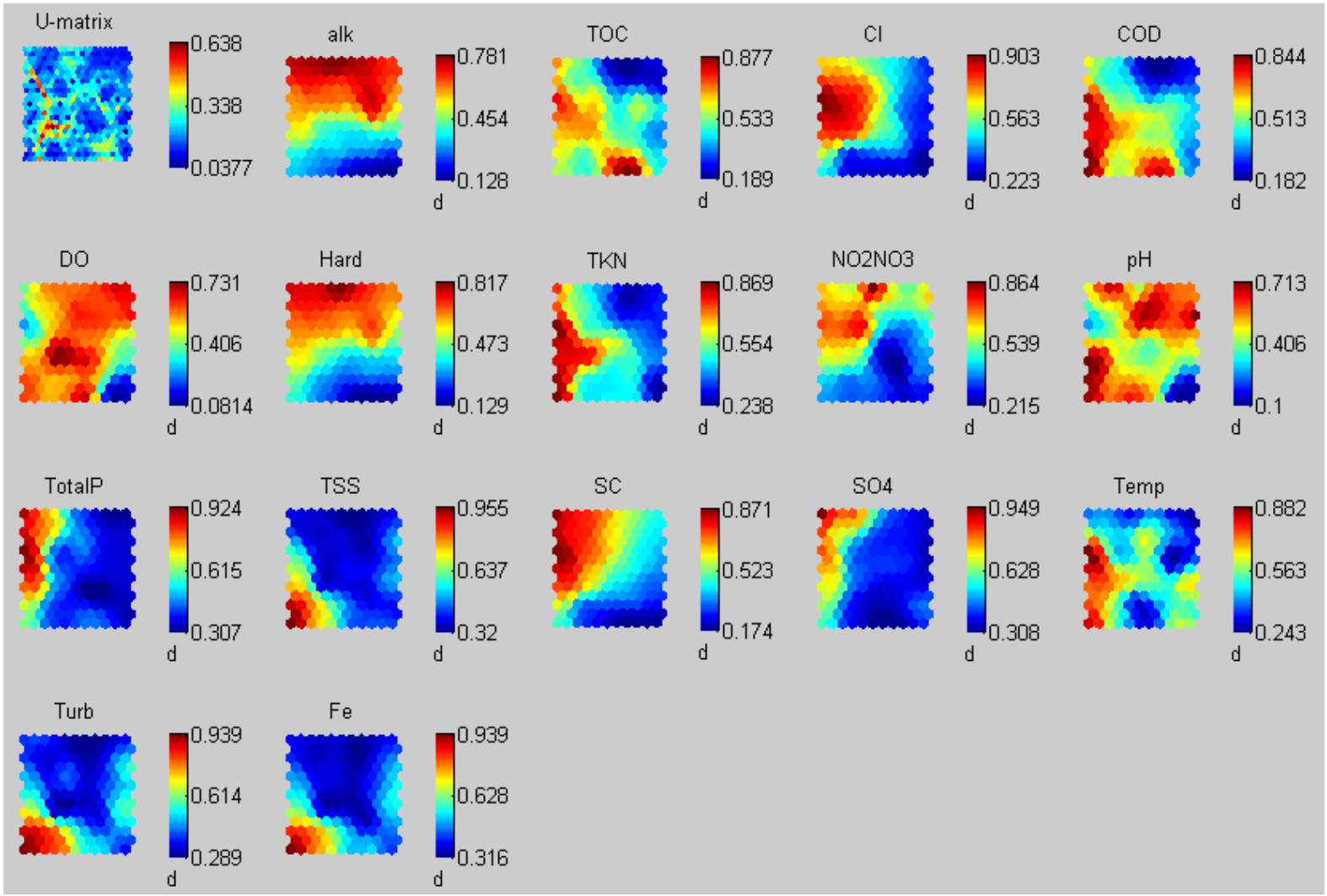
Supplementary Figure 1.14 – Quarter 3 Median Dataset Component Maps



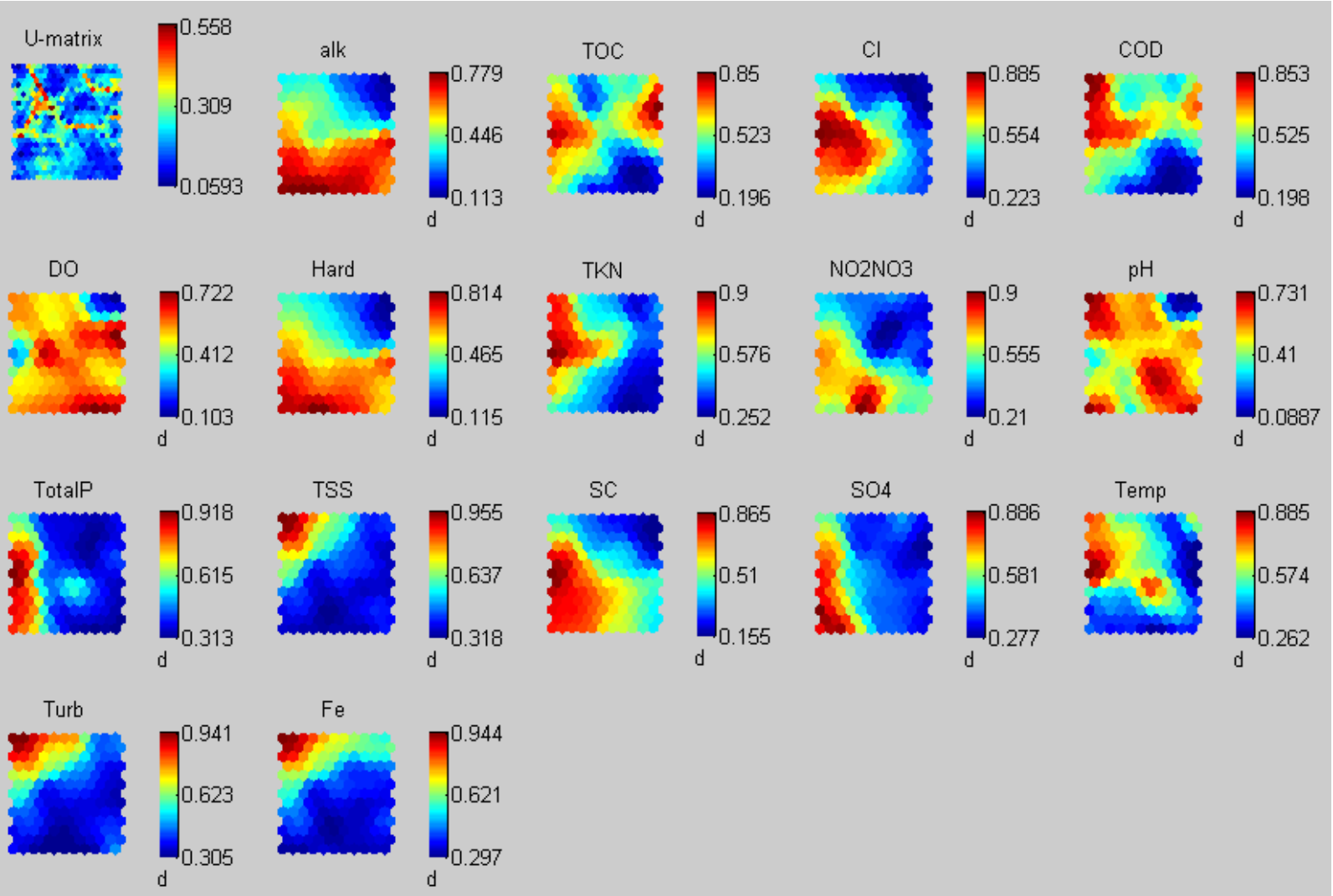
Supplementary Figure 1.15 – Quarter 3 Trimmed Mean Dataset Component Maps



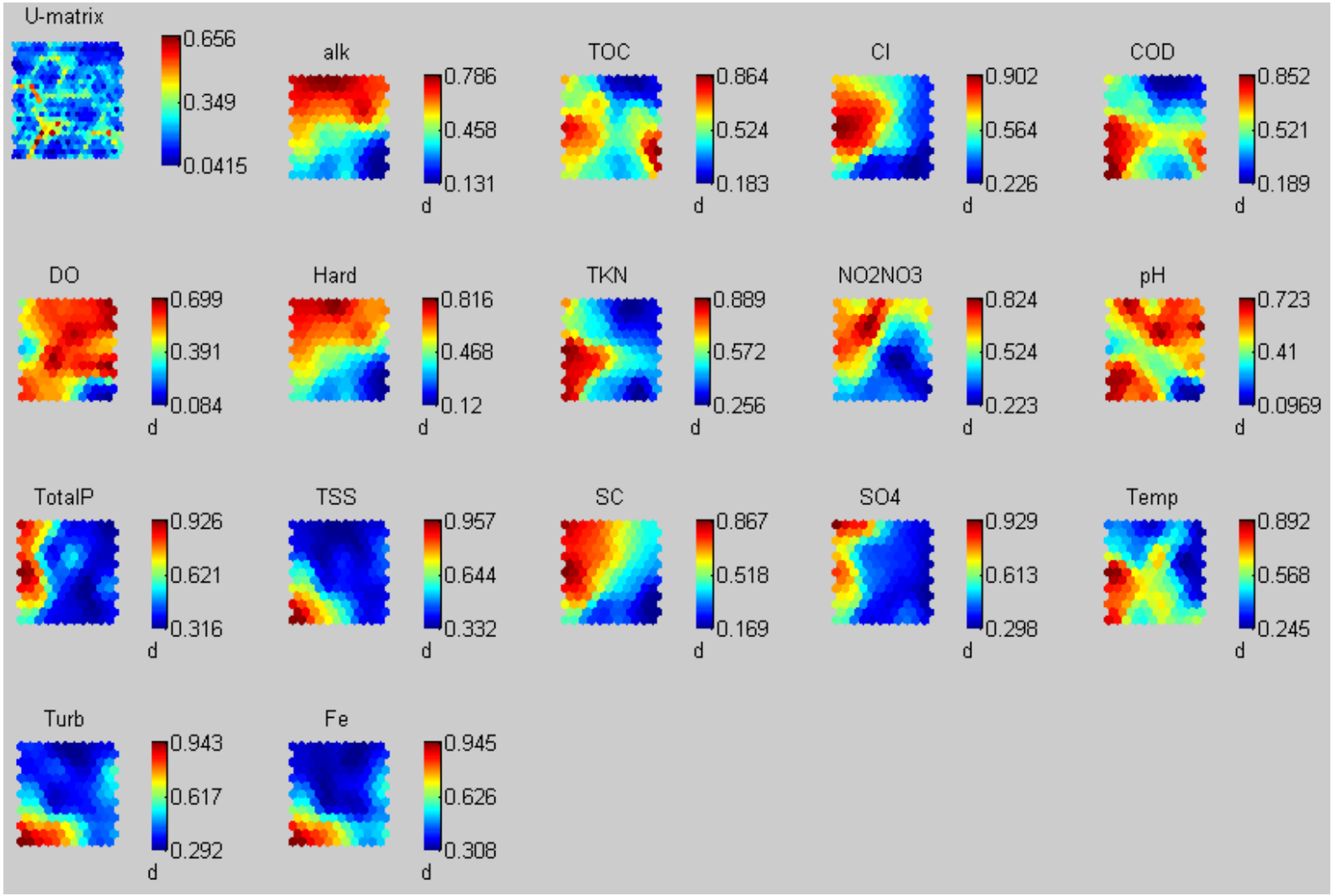
Supplementary Figure 1.16 – Quarter 3 Geometric Mean Dataset Component Maps



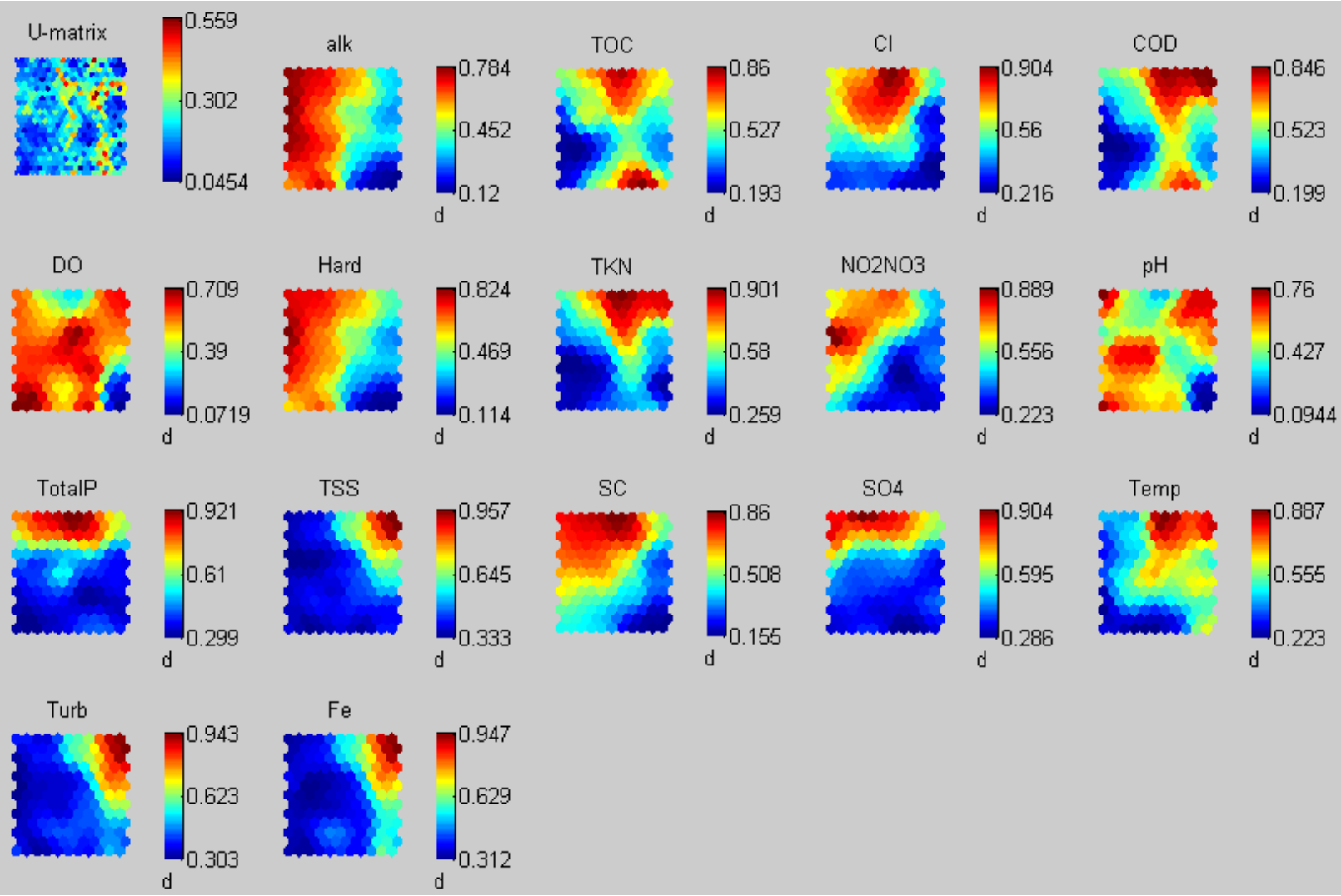
Supplementary Figure 1.17 – Quarter 4 Mean Dataset Component Maps



Supplementary Figure 1.18 – Quarter 4 Median Dataset Component Maps

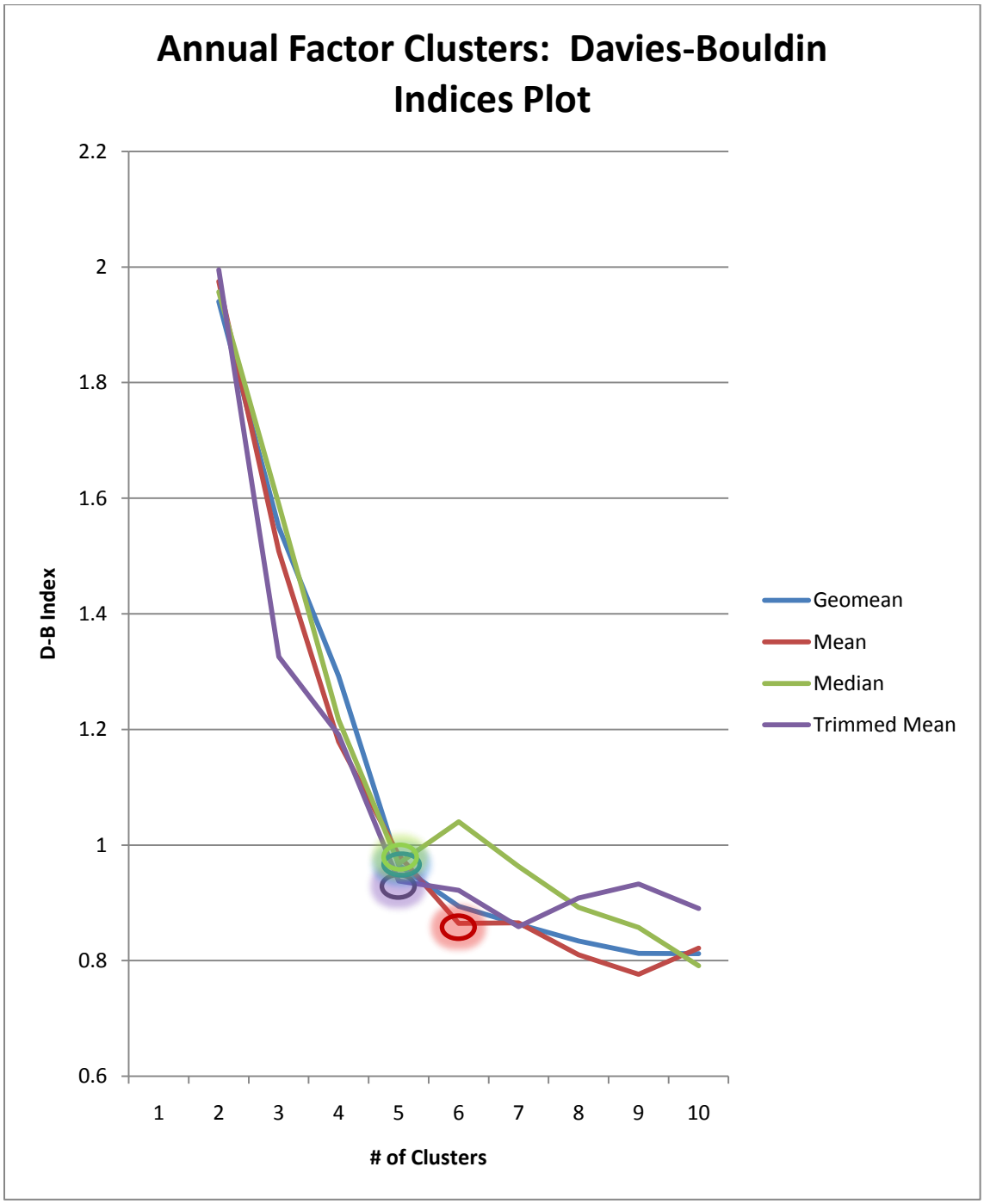


Supplementary Figure 1.19 – Quarter 4 Trimmed Mean Dataset Component Maps

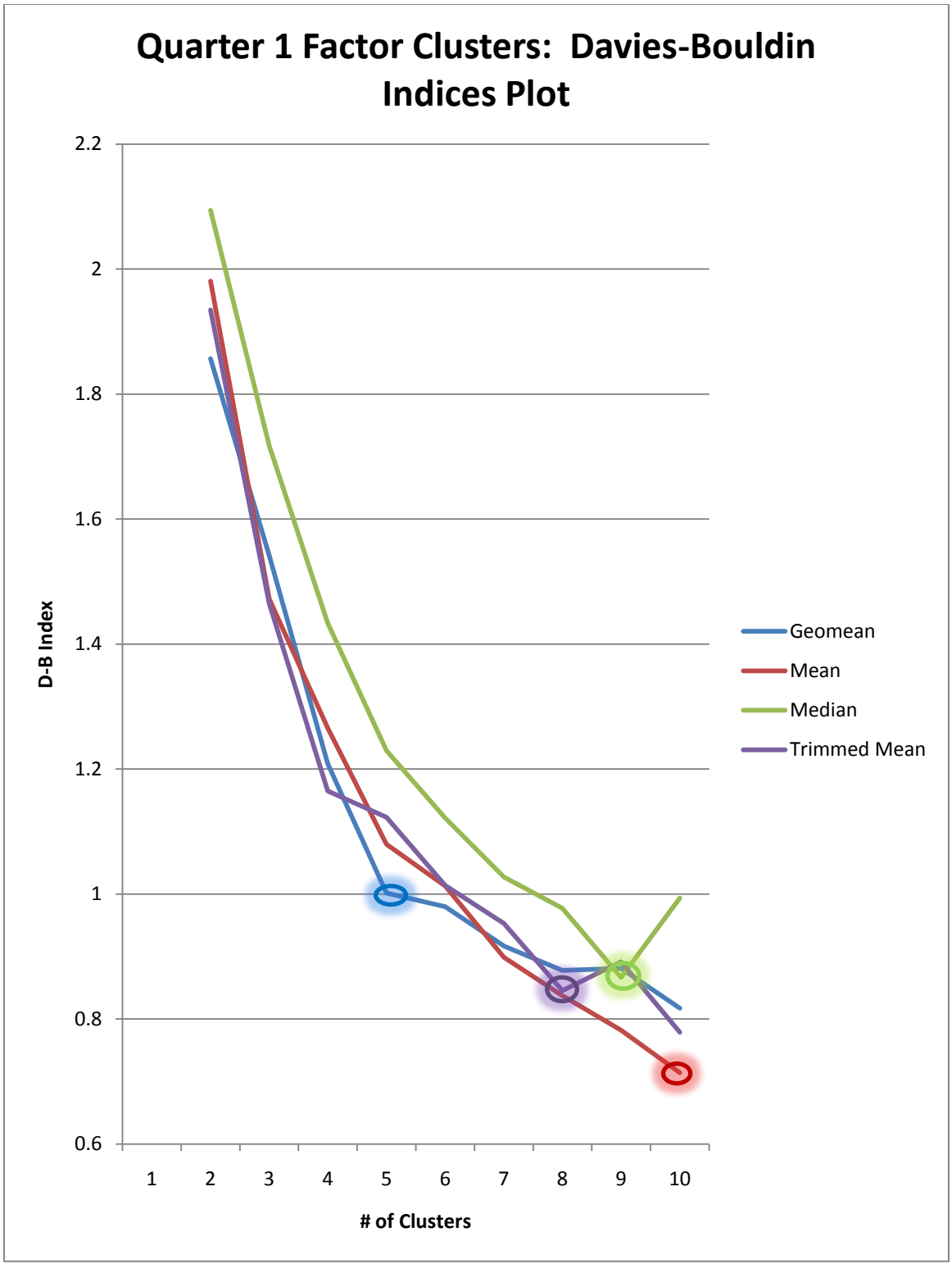


Supplementary Figure 1.20 – Quarter 4 Geometric Mean Dataset Component Maps

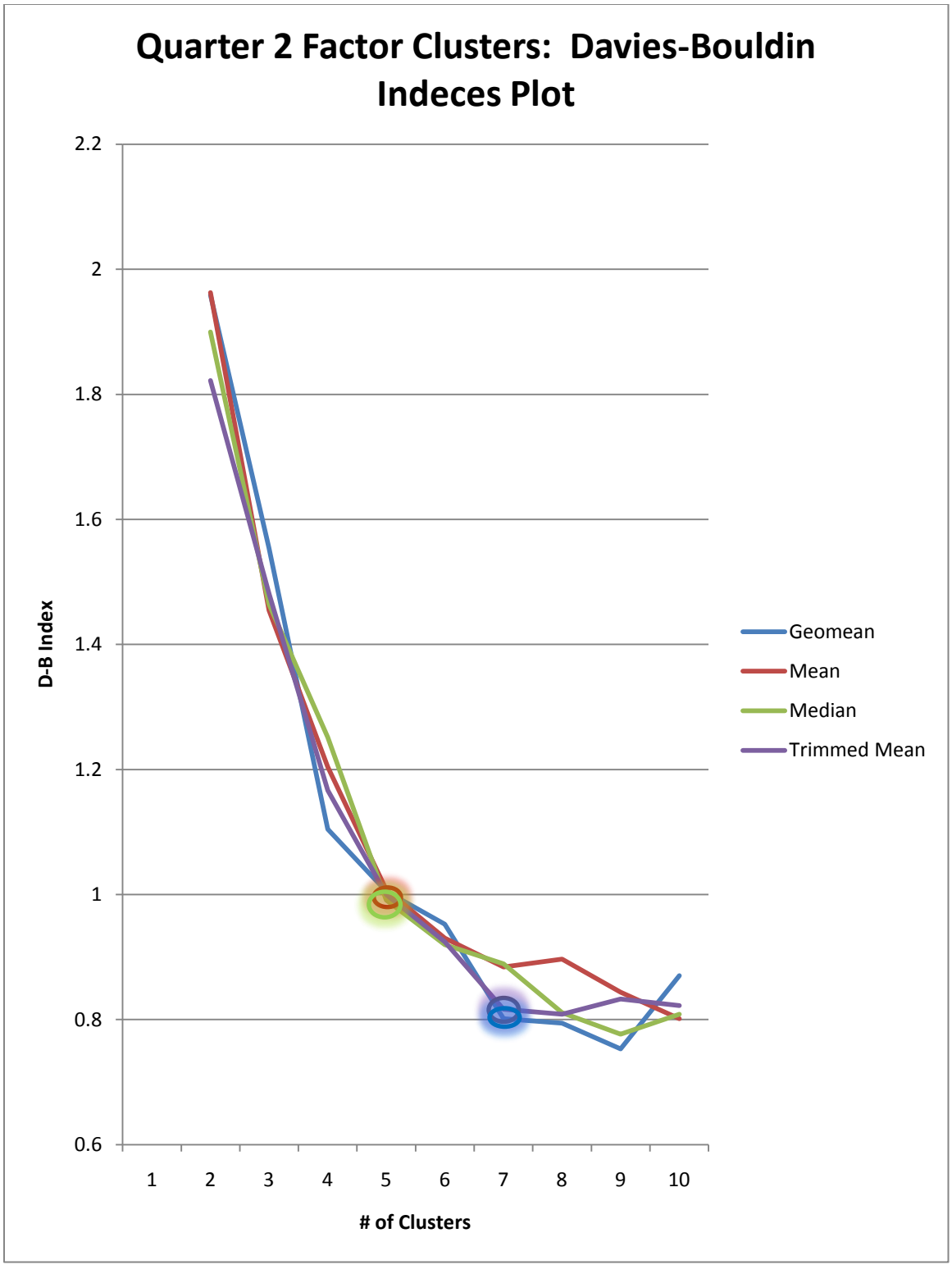
Davies-Bouldin Index Plots



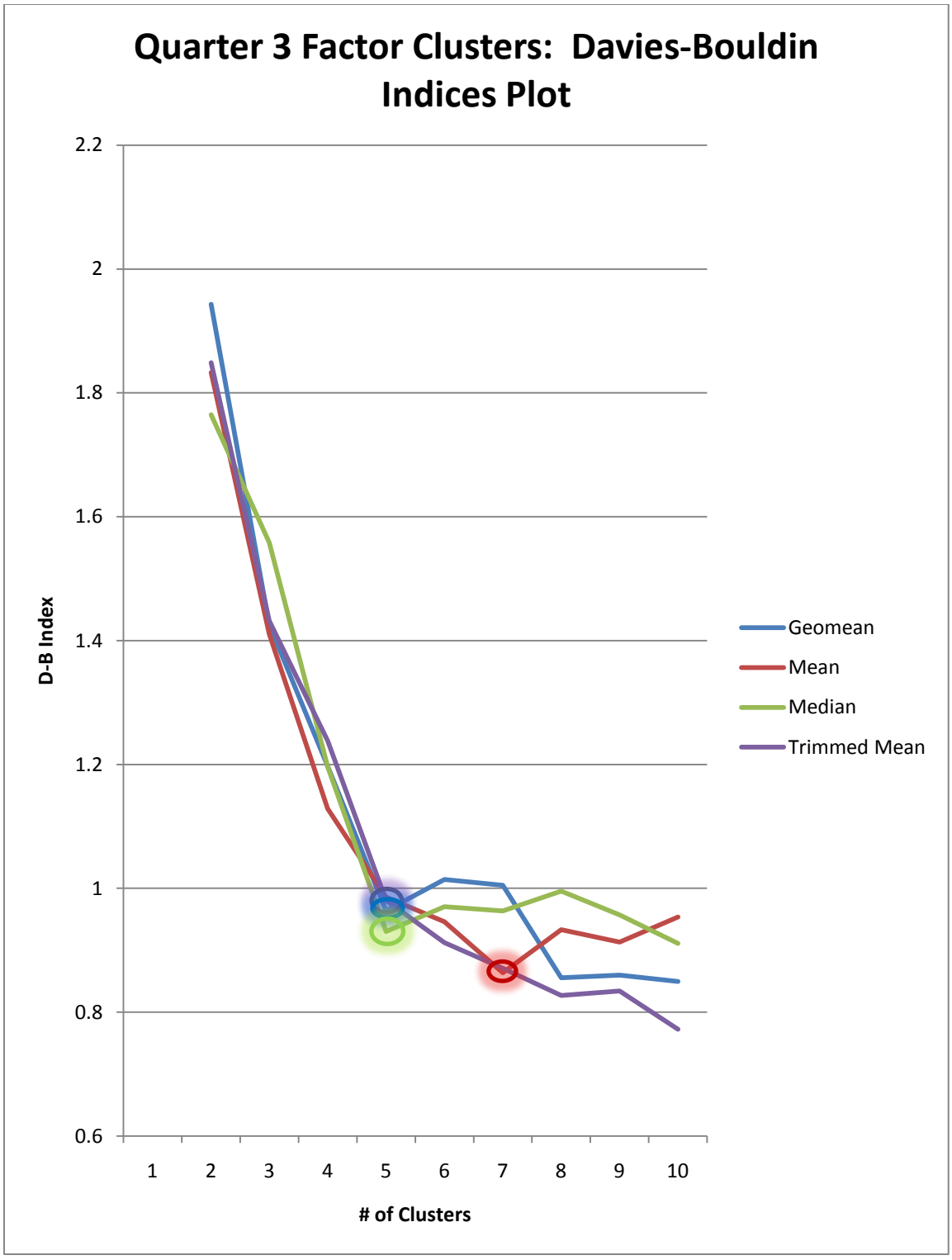
Supplementary Figure 2.1 – Davies-Bouldin indices plots for the cluster analysis of the annual factor datasets; five clusters were selected for the geometric mean dataset, six clusters for the mean dataset, five clusters for the median dataset, and five clusters for the trimmed mean dataset



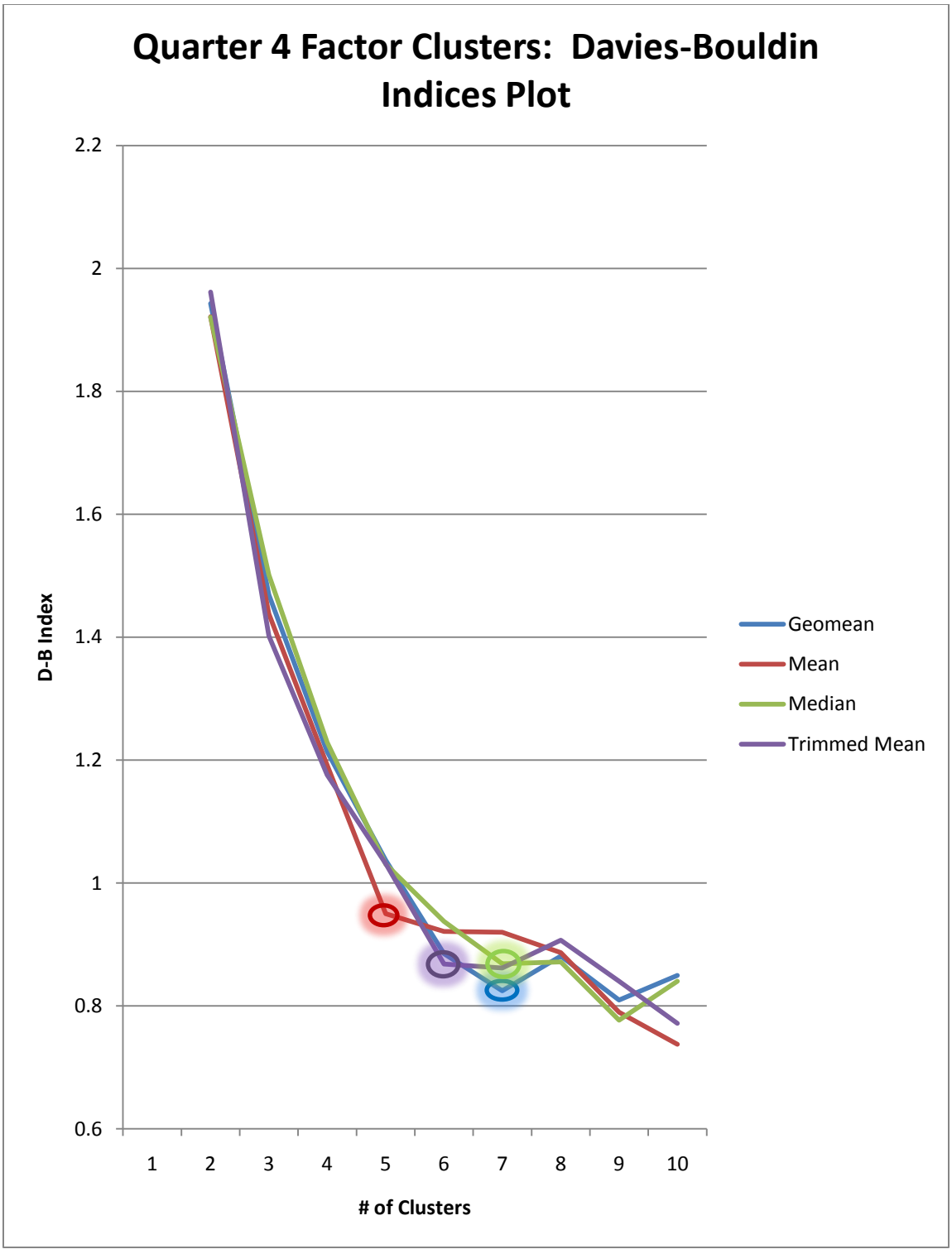
Supplementary Figure 2.2 – Davies-Bouldin indices plots for the cluster analysis of the quarter 1 factor datasets five clusters were selected for the geometric mean dataset, ten clusters for the mean dataset, nine clusters for the median dataset, and eight clusters for the trimmed mean dataset



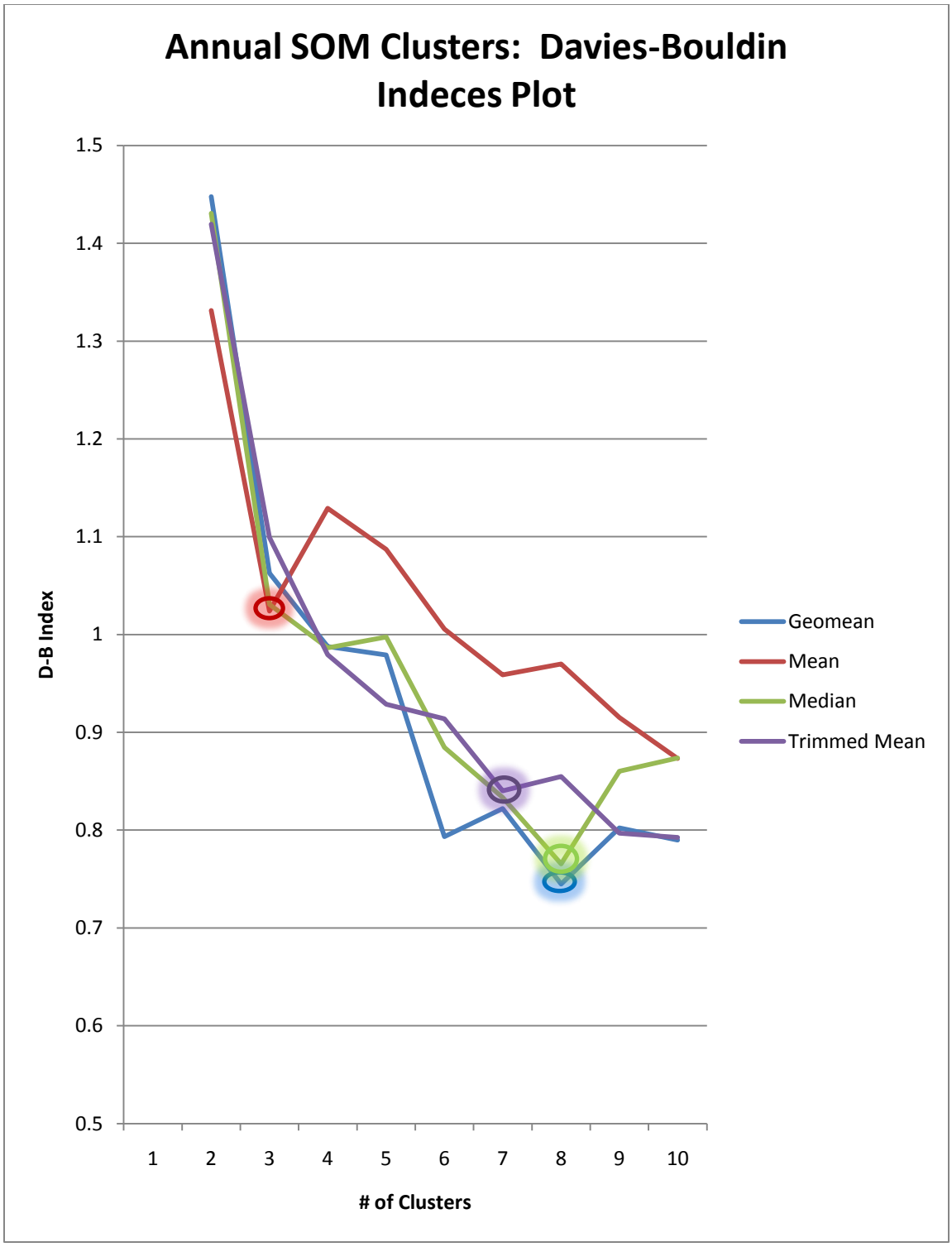
Supplementary Figure 2.3 – Davies-Bouldin indices plots for the cluster analysis of the quarter 2 factor datasets; seven clusters were selected for the geometric mean dataset, five clusters for the mean dataset, five clusters for the median dataset, and seven clusters for the trimmed mean dataset



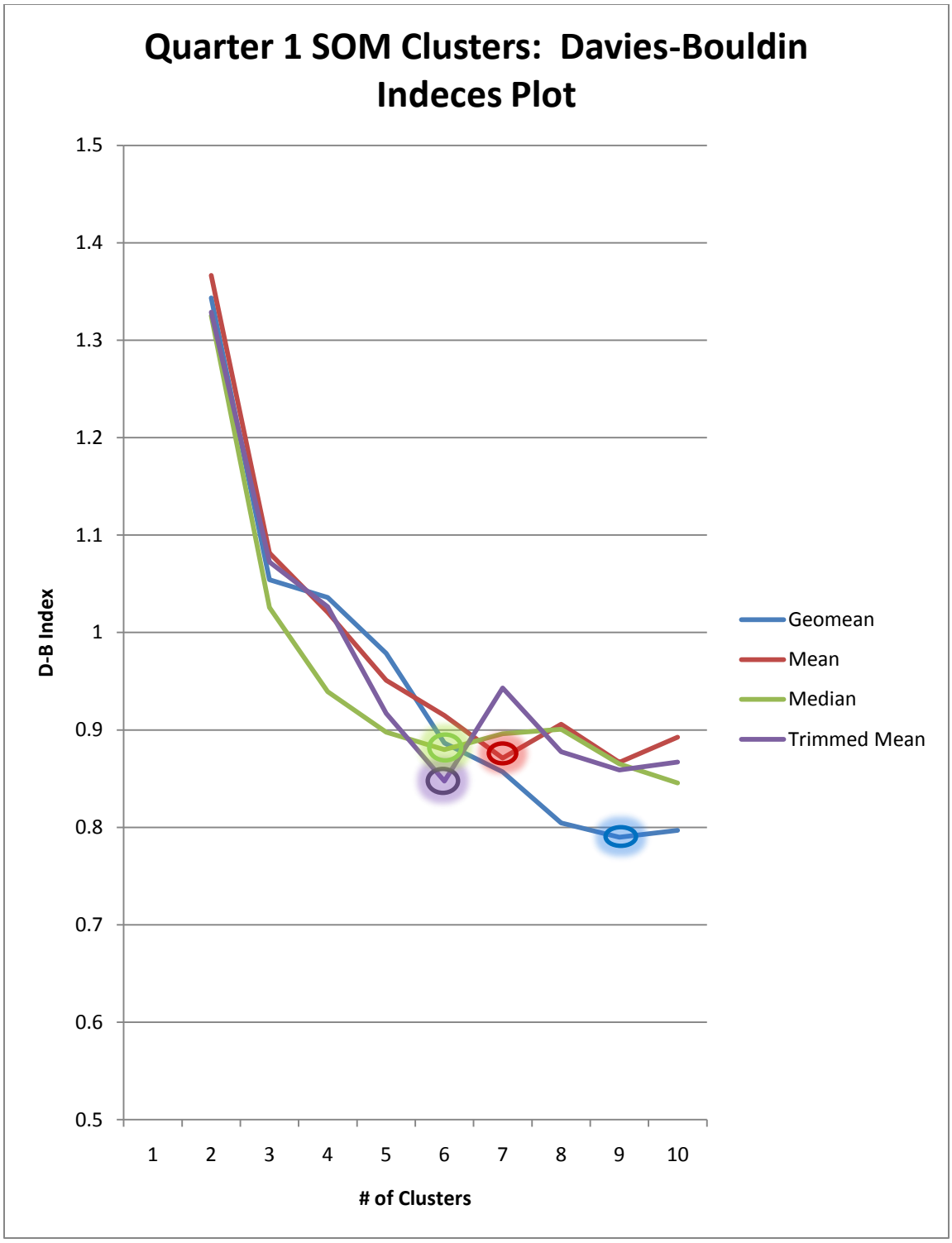
Supplementary Figure 2.4 – Davies-Bouldin indices plots for the cluster analysis of the quarter 3 factor datasets; five clusters were selected for the geometric mean dataset, seven clusters for the mean dataset, five clusters for the median dataset, and five clusters for the trimmed mean dataset



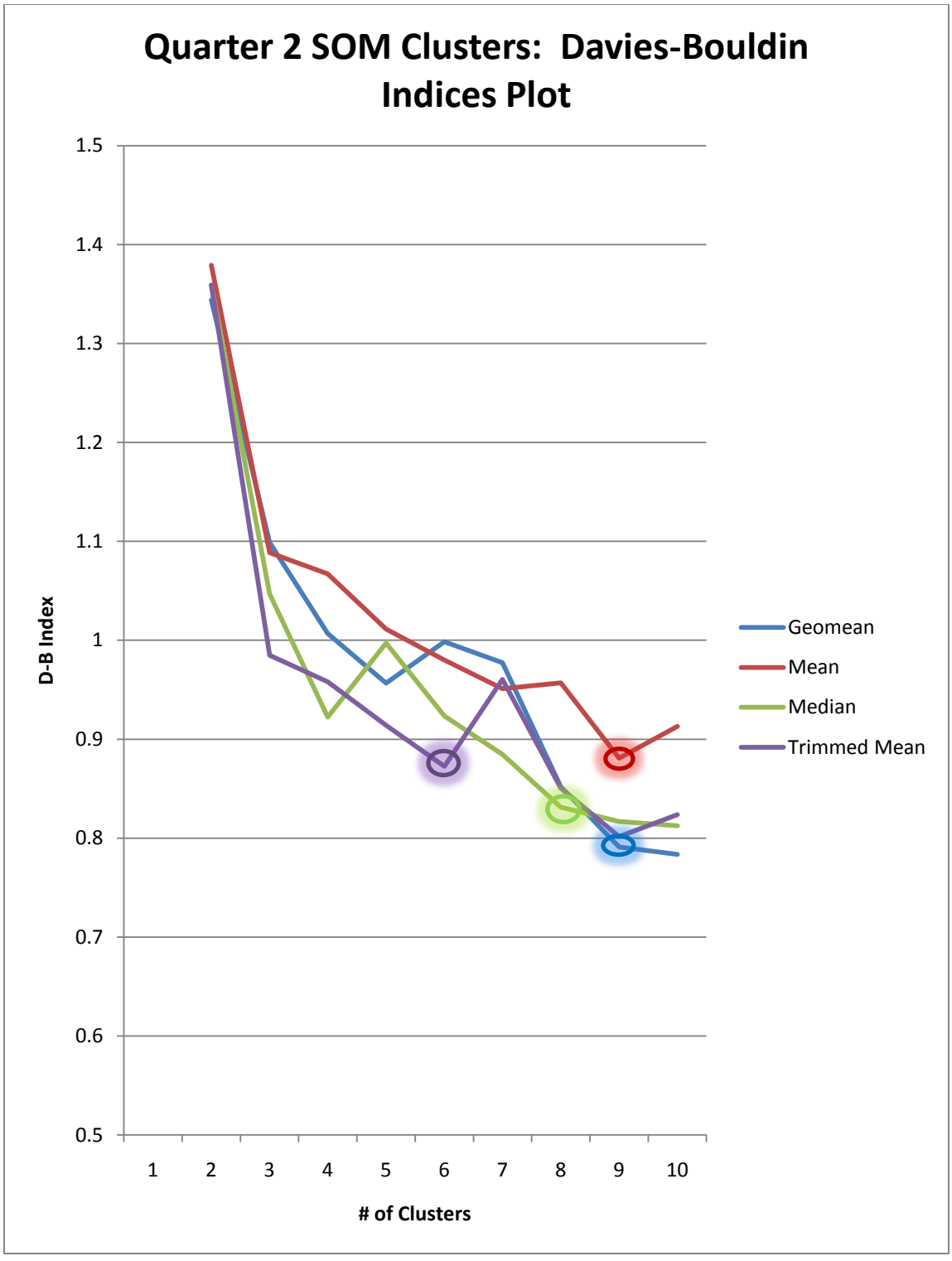
Supplementary Figure 2.5 – Davies-Bouldin indices plots for the cluster analysis of the quarter 4 factor datasets; seven clusters were selected for the geometric mean dataset, five clusters for the mean dataset, seven clusters for the median dataset, and six clusters for the trimmed mean dataset



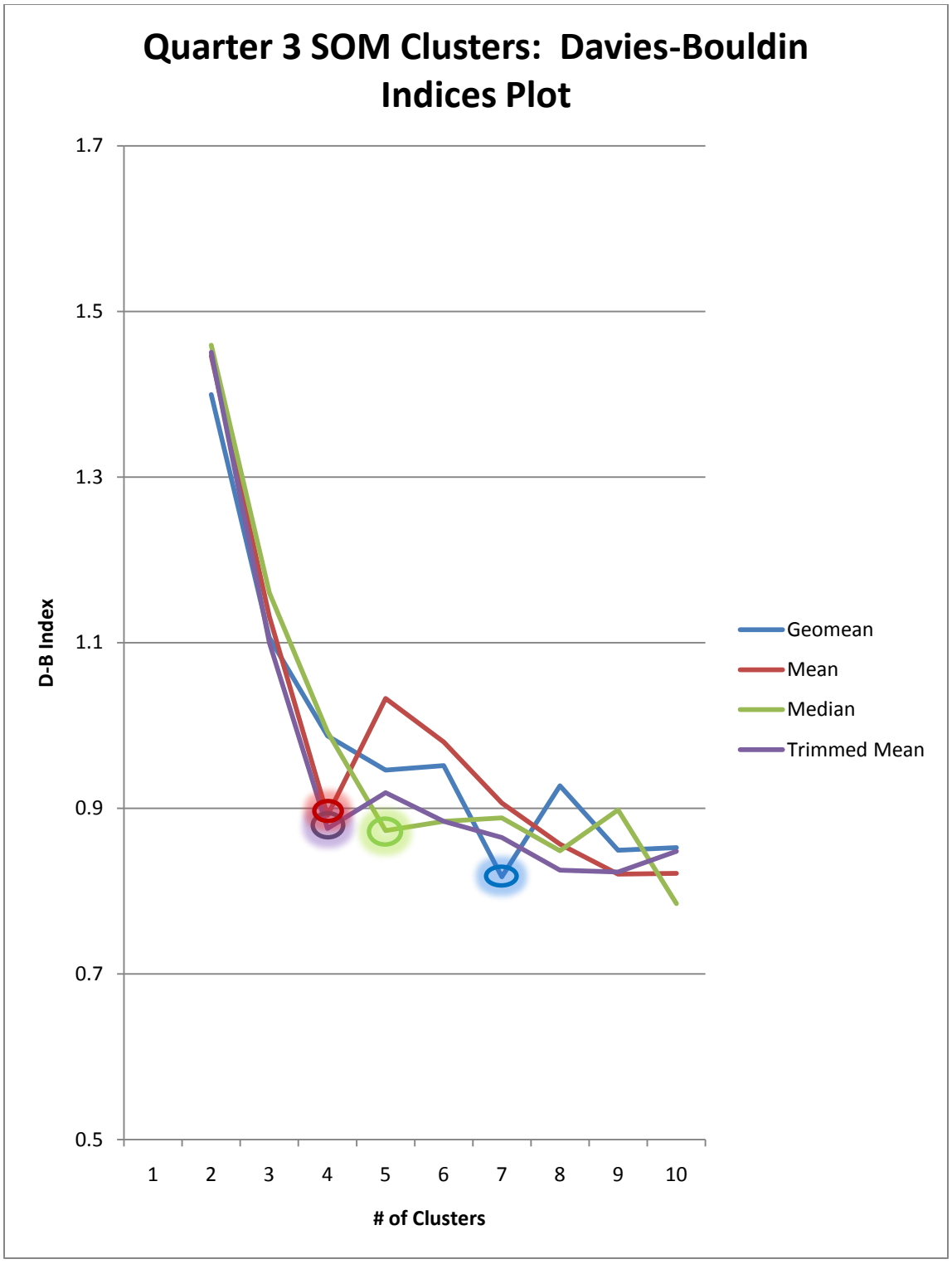
Supplementary Figure 2.6 – Davies-Bouldin indices plots for the cluster analysis of the annual SOM datasets; eight clusters were selected for the geometric mean dataset, three clusters for the mean dataset, eight clusters for the median dataset, and seven clusters for the trimmed mean dataset



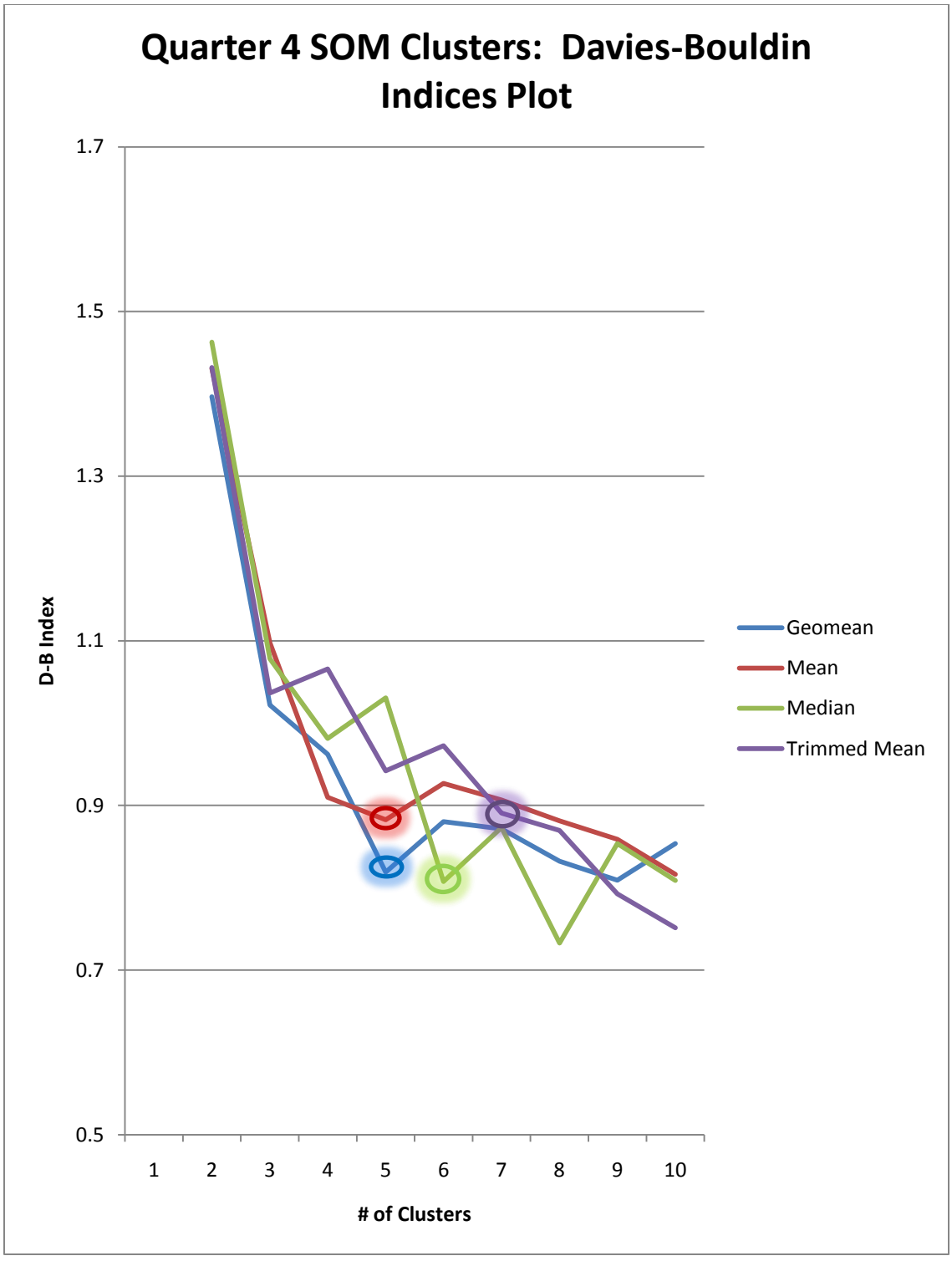
Supplementary Figure 2.7 – Davies-Bouldin indices plot for the cluster analysis of the quarter 1 SOM datasets; nine clusters were selected for the geometric mean dataset, seven clusters for the mean dataset, six clusters for the median dataset, and six clusters for the trimmed mean dataset



Supplementary Figure 2.8 – Davies-Bouldin indices plot for the cluster analysis of the quarter 2 SOM datasets; nine clusters were selected for the geometric mean dataset, nine clusters for the mean dataset, eight clusters for the median dataset, and six clusters for the trimmed mean dataset

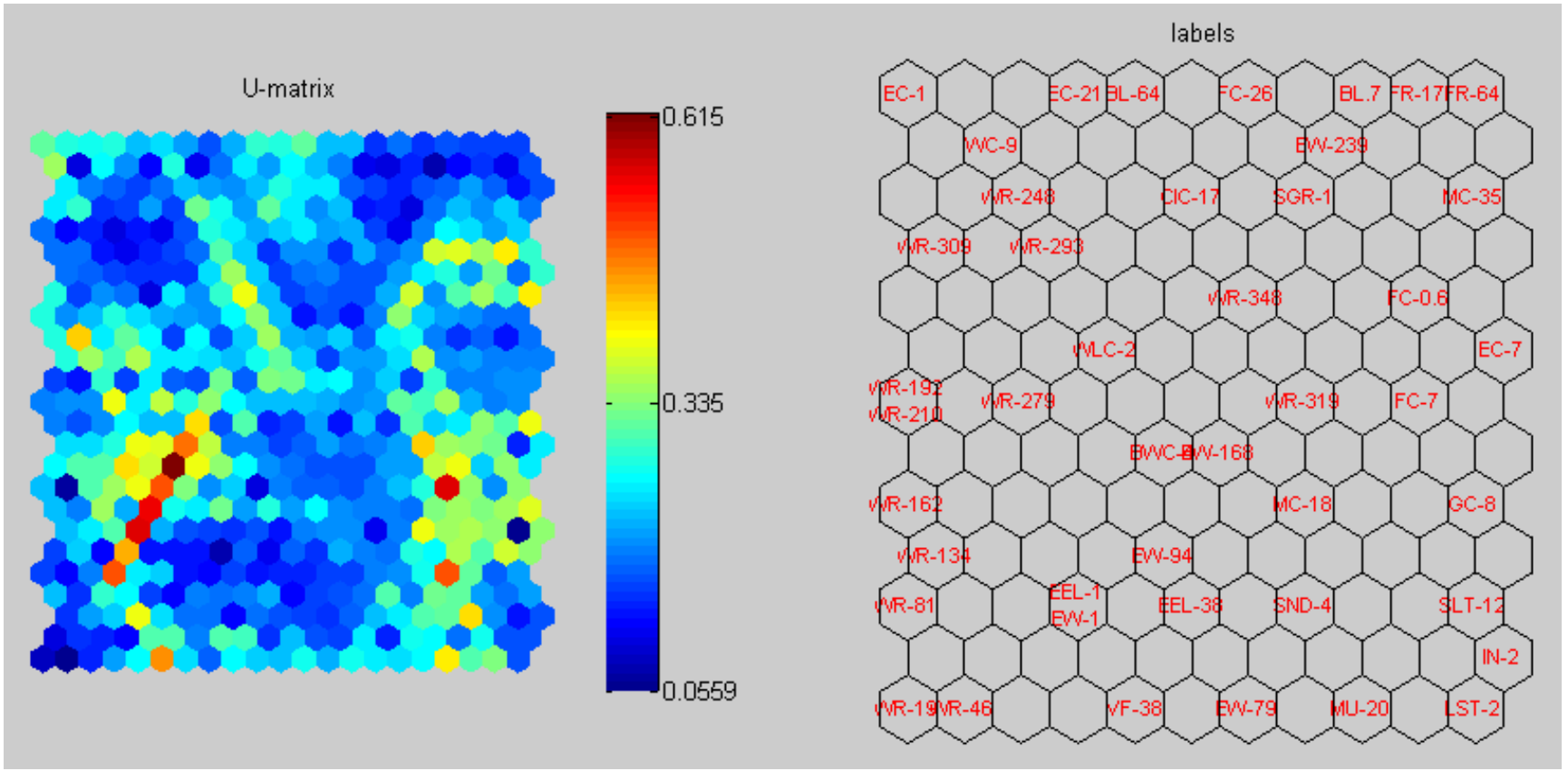


Supplementary Figure 2.9 – Davies-Bouldin indices plot for the cluster analysis of the quarter 3 SOM datasets; seven clusters were selected for the geometric mean dataset, four clusters for the mean dataset, five clusters for the median dataset, and four clusters for the trimmed mean dataset

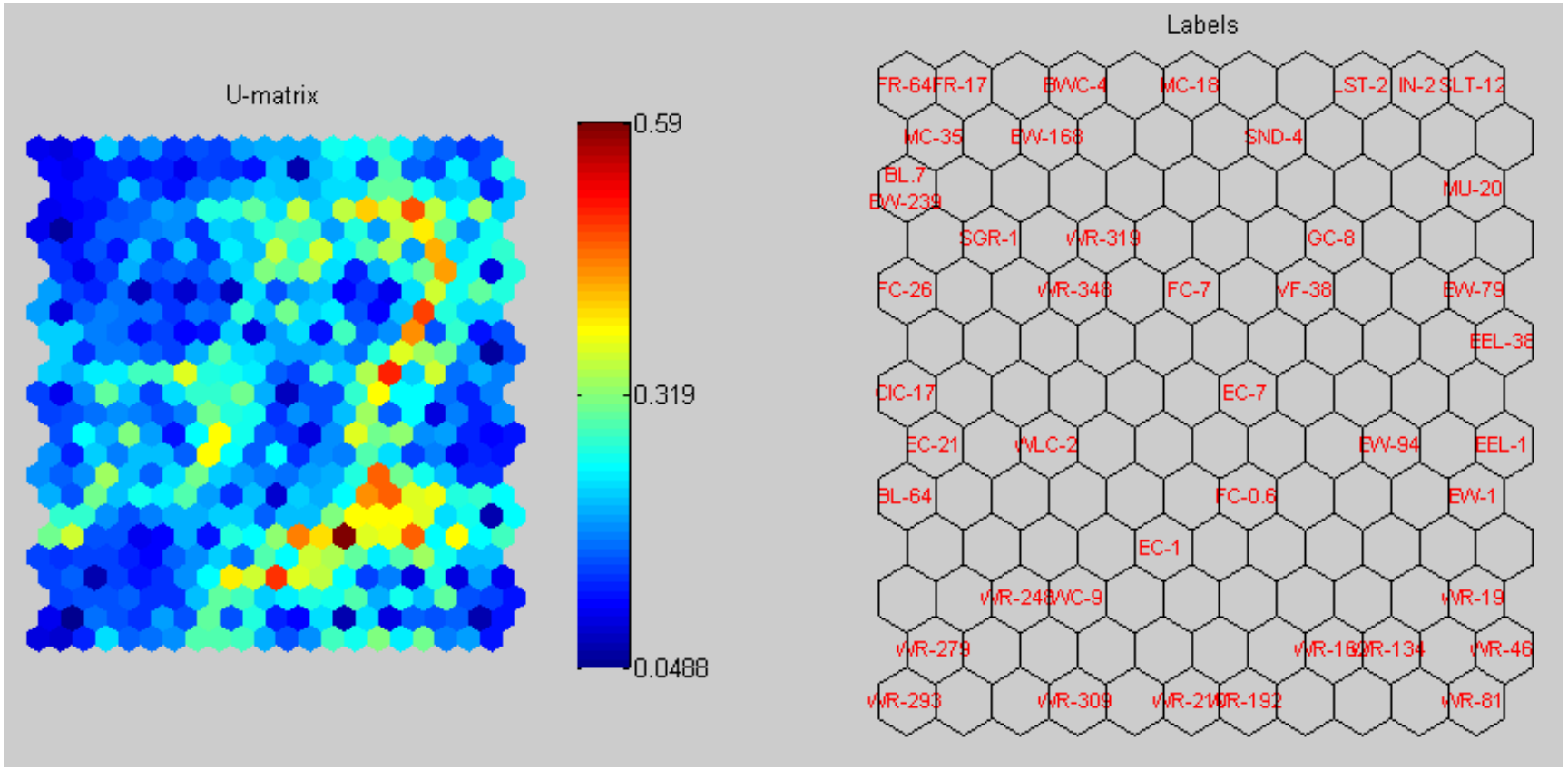


Supplementary Figure 2.10 – Davies-Bouldin indices plot for the cluster analysis of the quarter 4 SOM datasets; five clusters were selected for the geometric mean dataset, five clusters for the mean dataset, six clusters for the median dataset, and seven clusters for the trimmed mean dataset

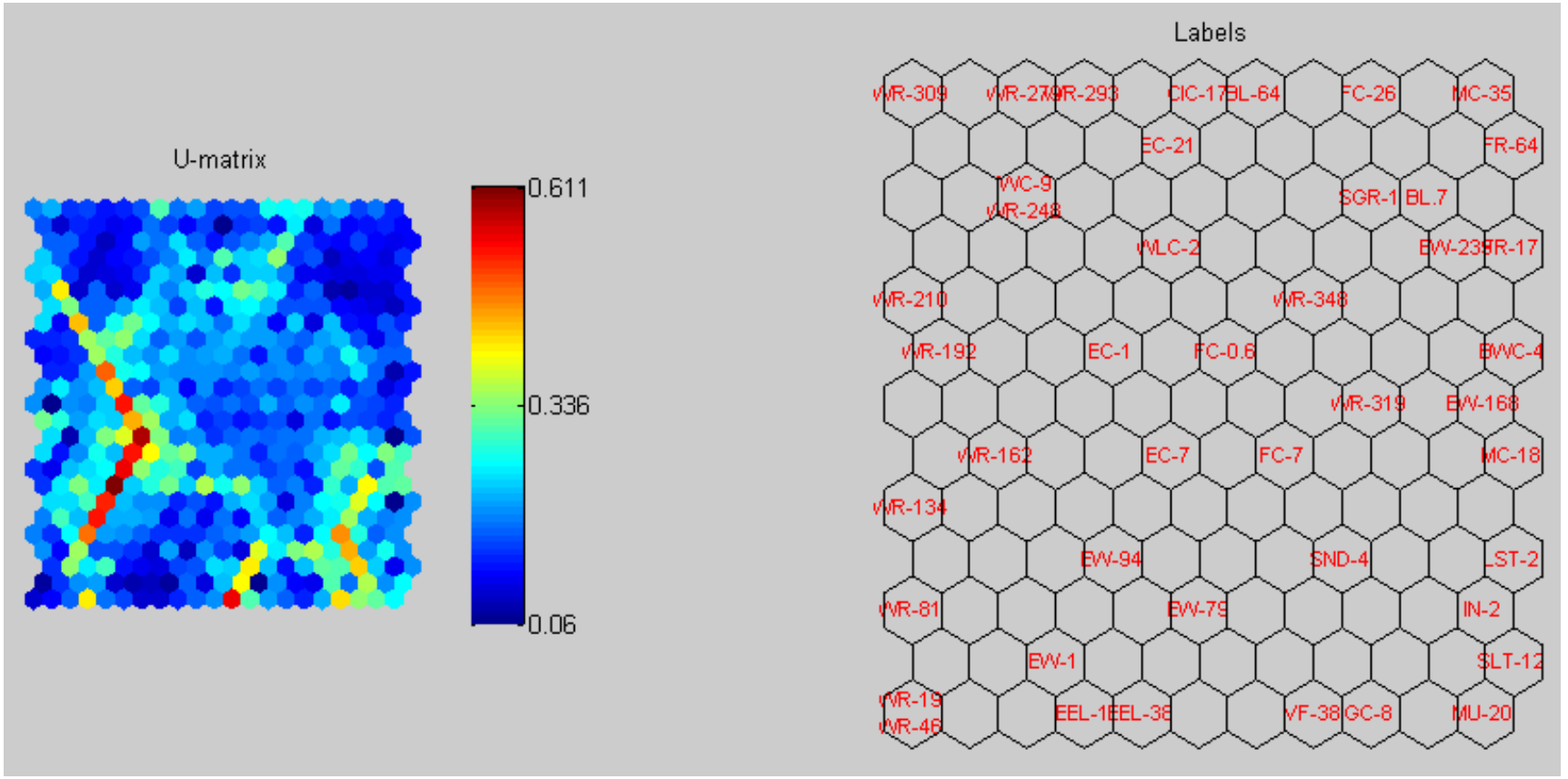
SOM Unified Distance Matrices



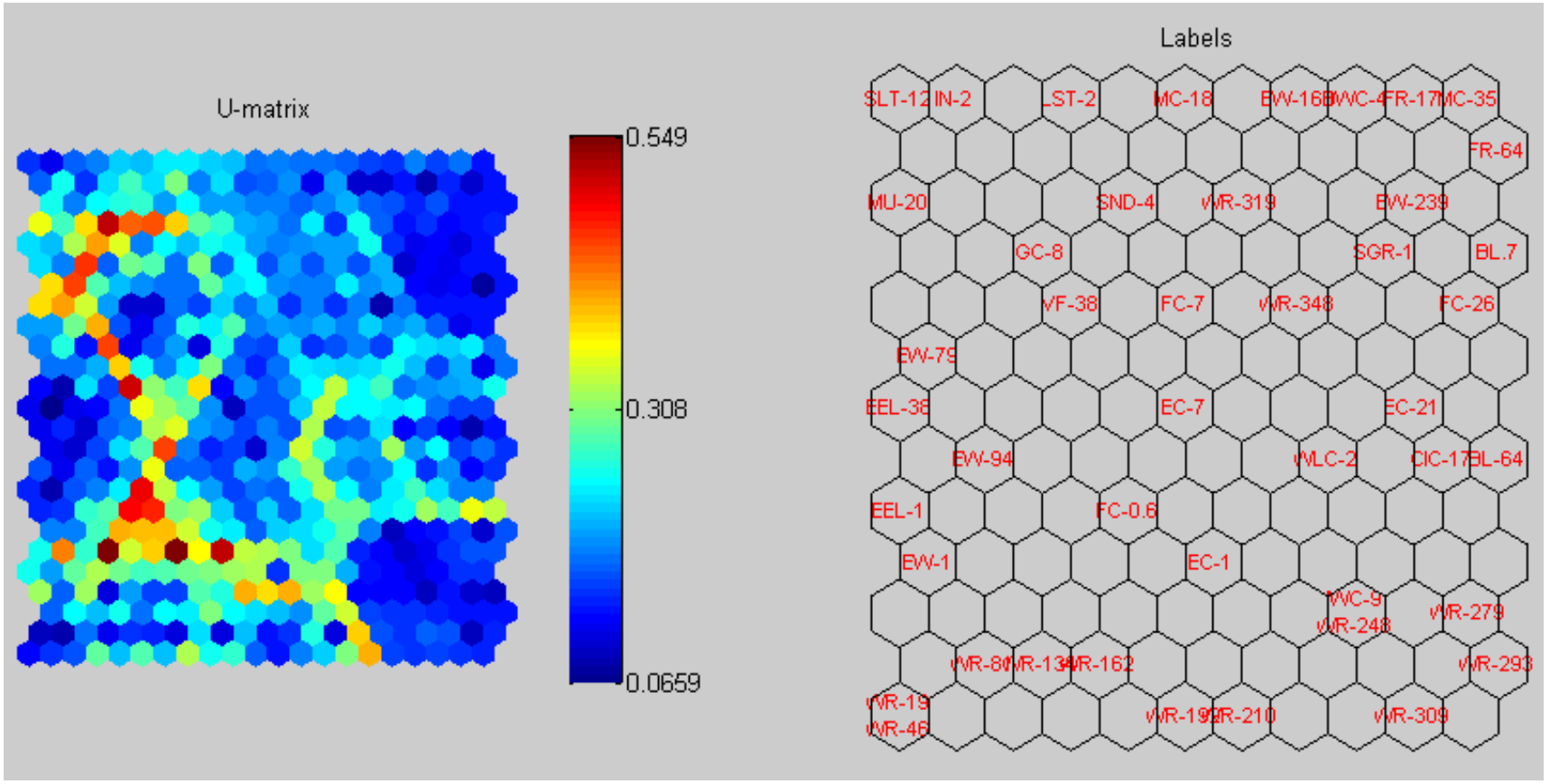
Supplementary Figure 3.1 – Annual Mean dataset U-Matrix and station organization on the SOM



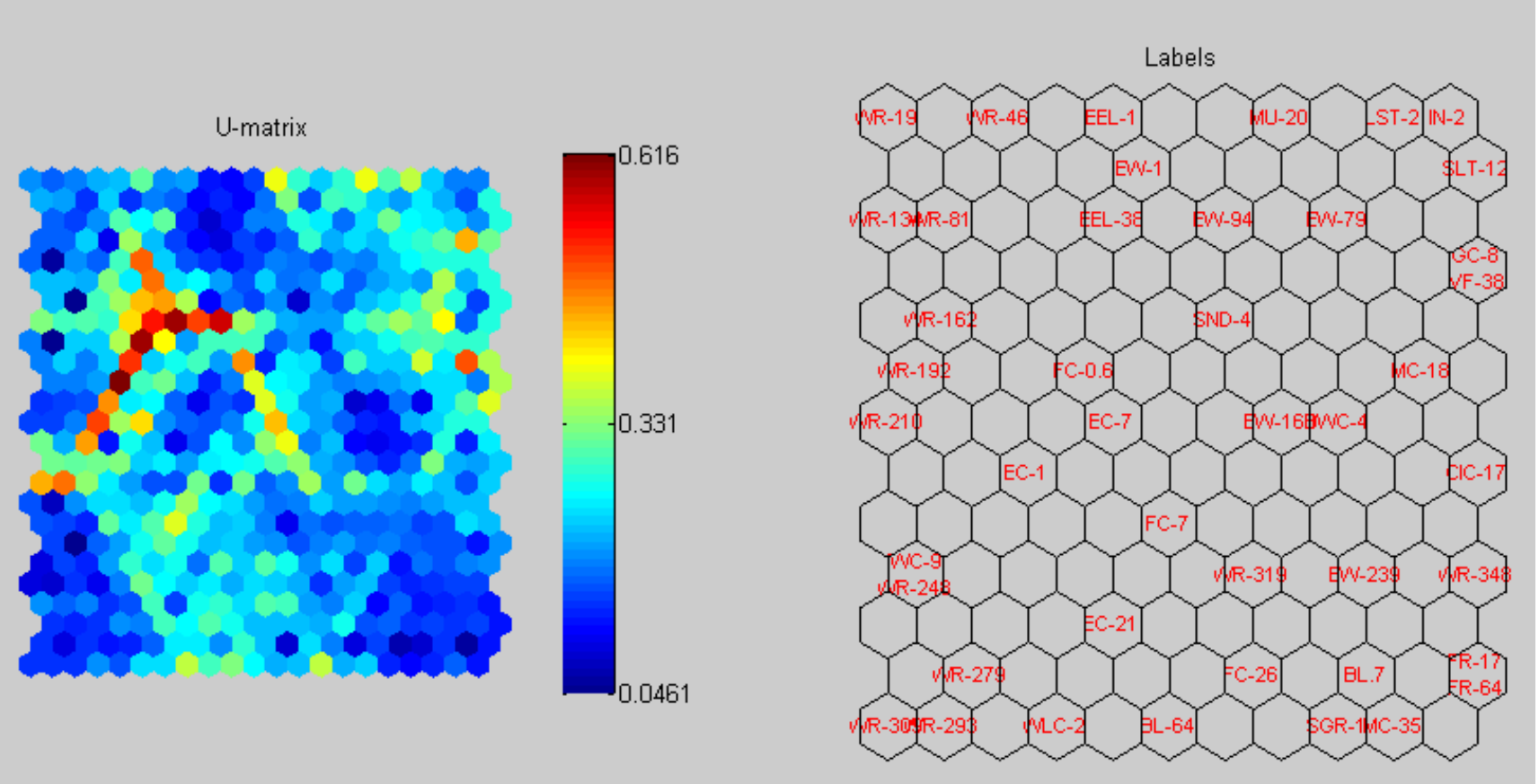
Supplementary Figure 3.2 – Annual Median dataset U-Matrix and station organization on the SOM



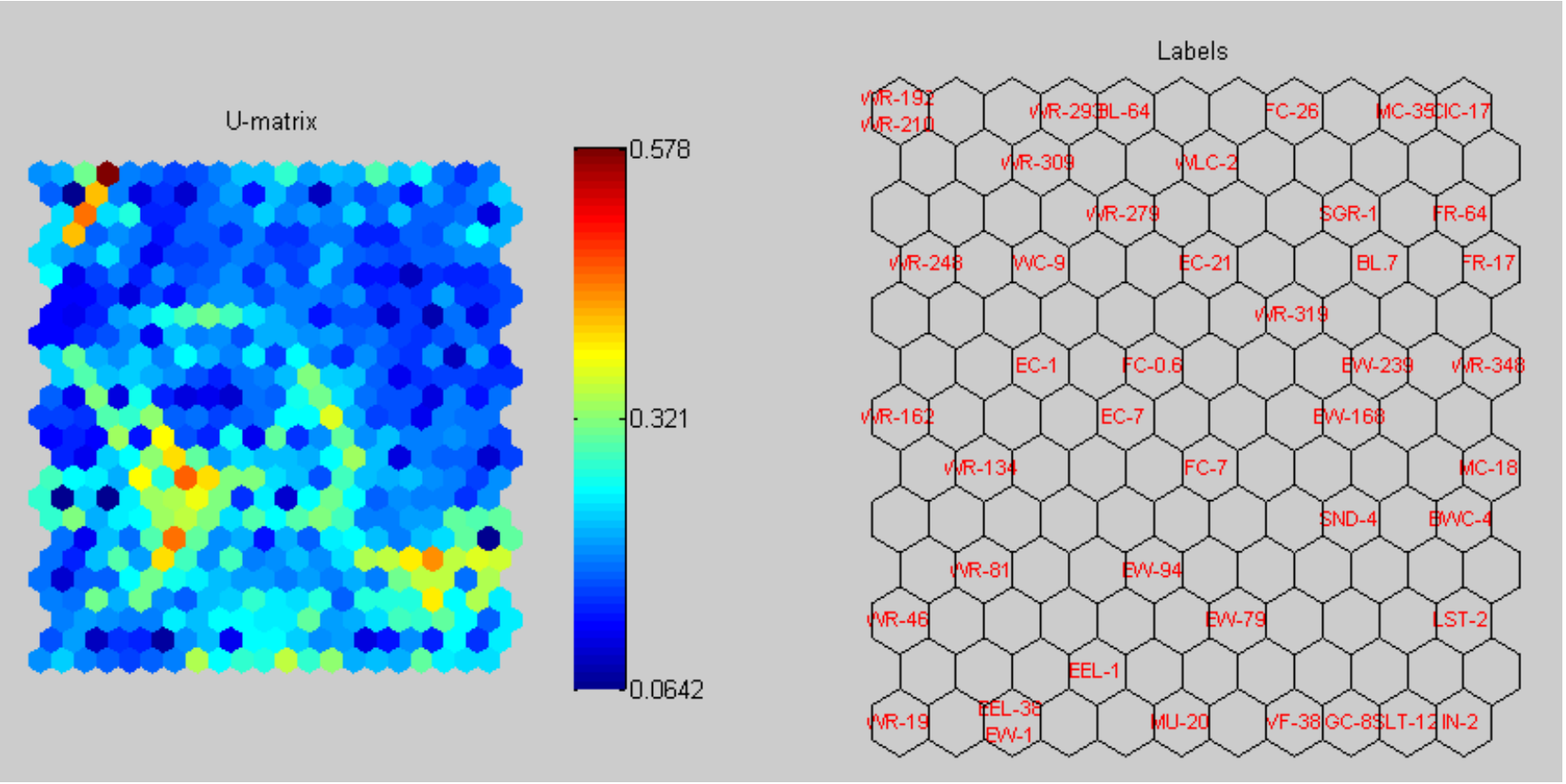
Supplementary Figure 3.3 – Annual Trimmed Mean dataset U-Matrix and station organization on the SOM



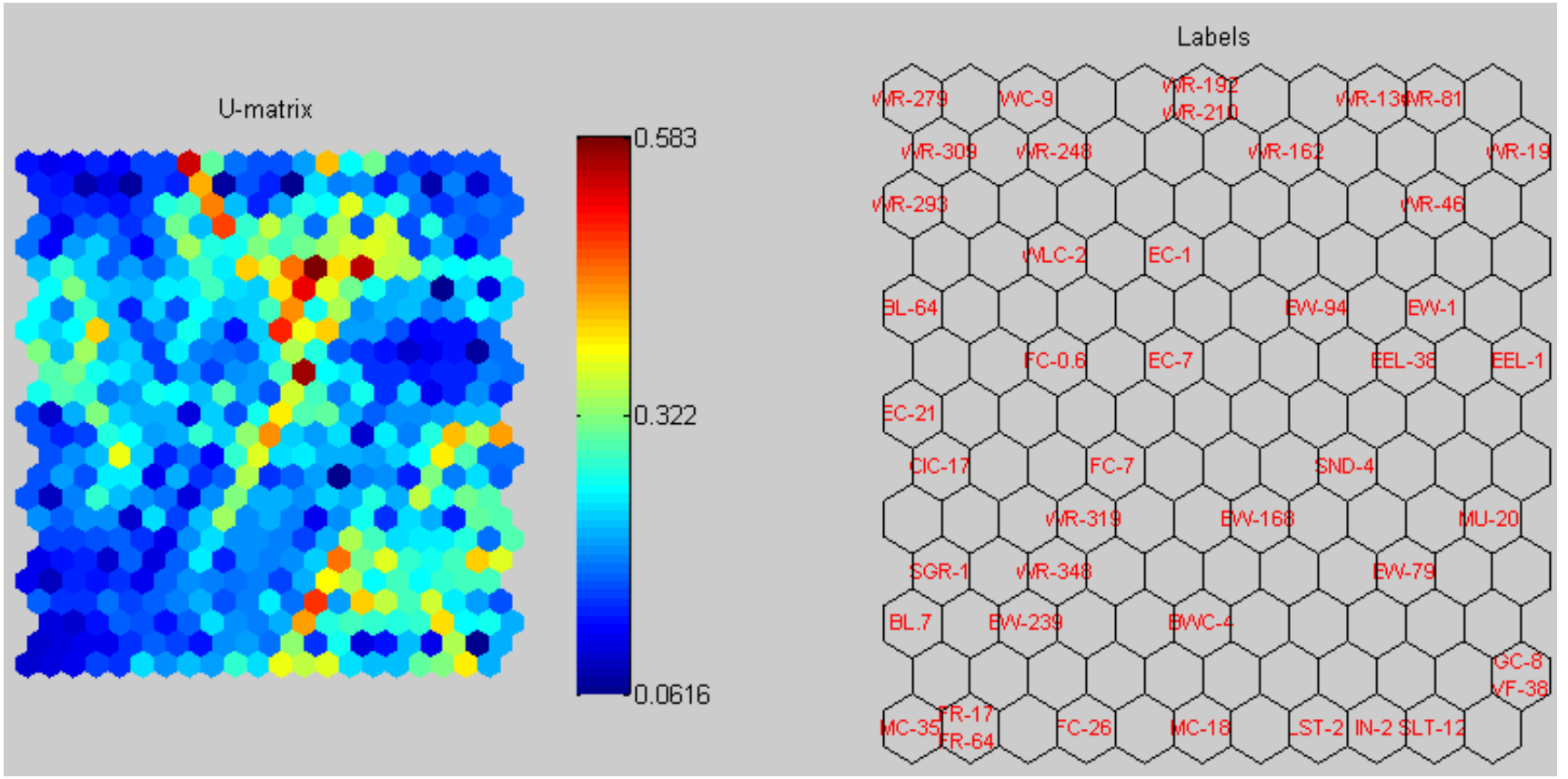
Supplementary Figure 3.4 – Annual Geometric Mean dataset U-Matrix and station organization on the SOM



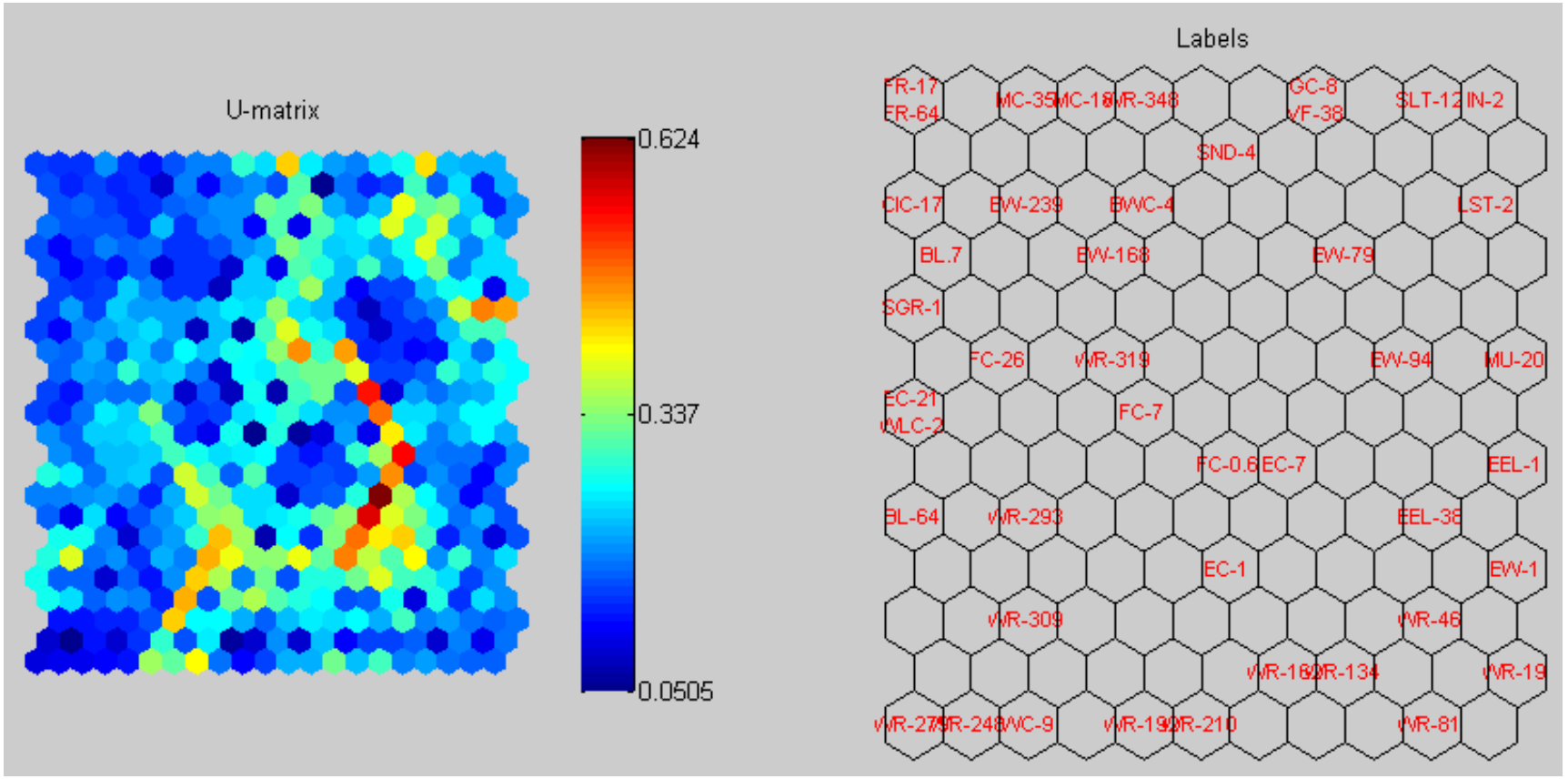
Supplementary Figure 3.5 – Quarter 1 Mean dataset U-Matrix and station organization on the SOM



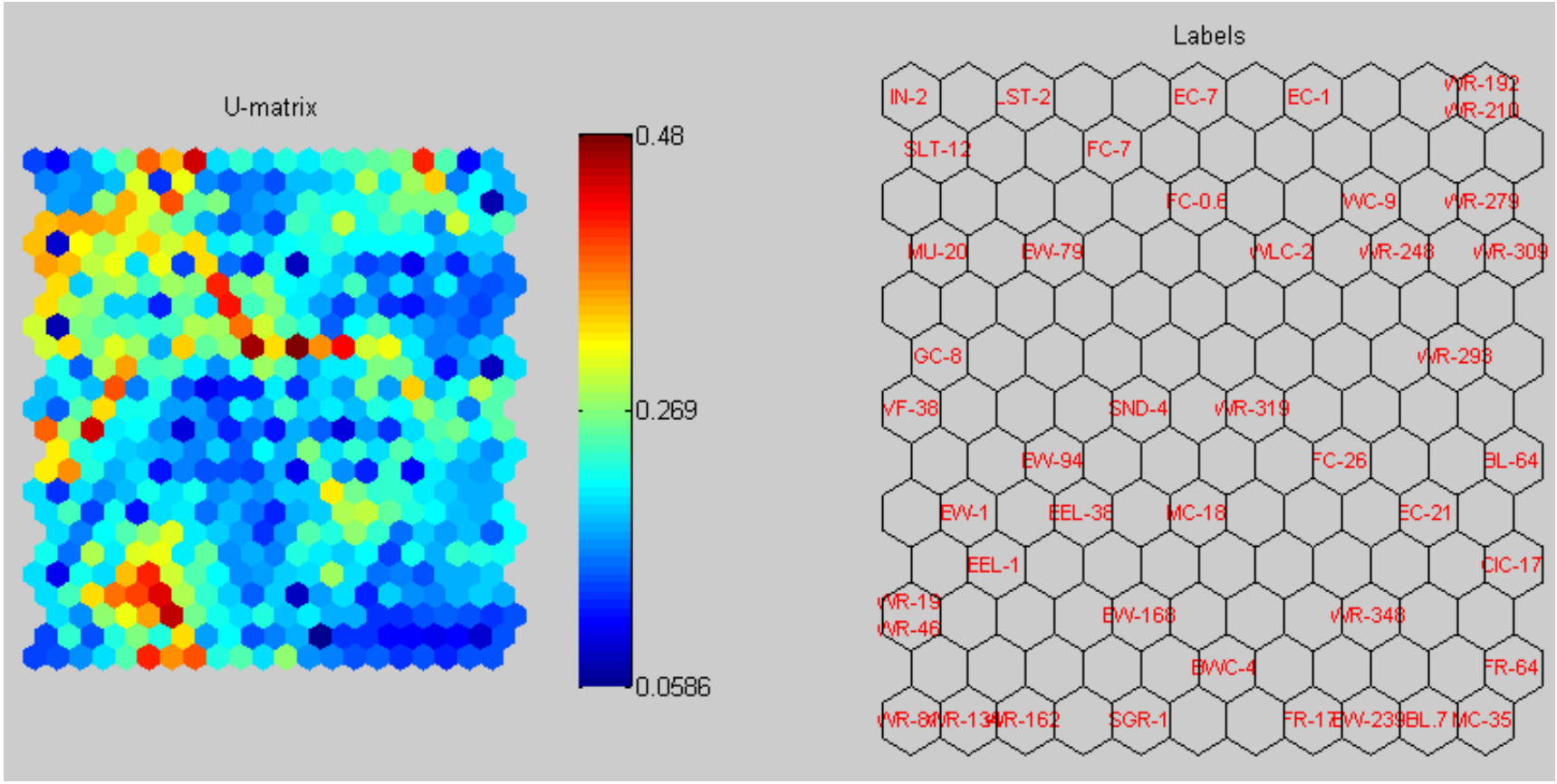
Supplementary Figure 3.6 – Quarter 1 Median dataset U-Matrix and station organization on the SOM



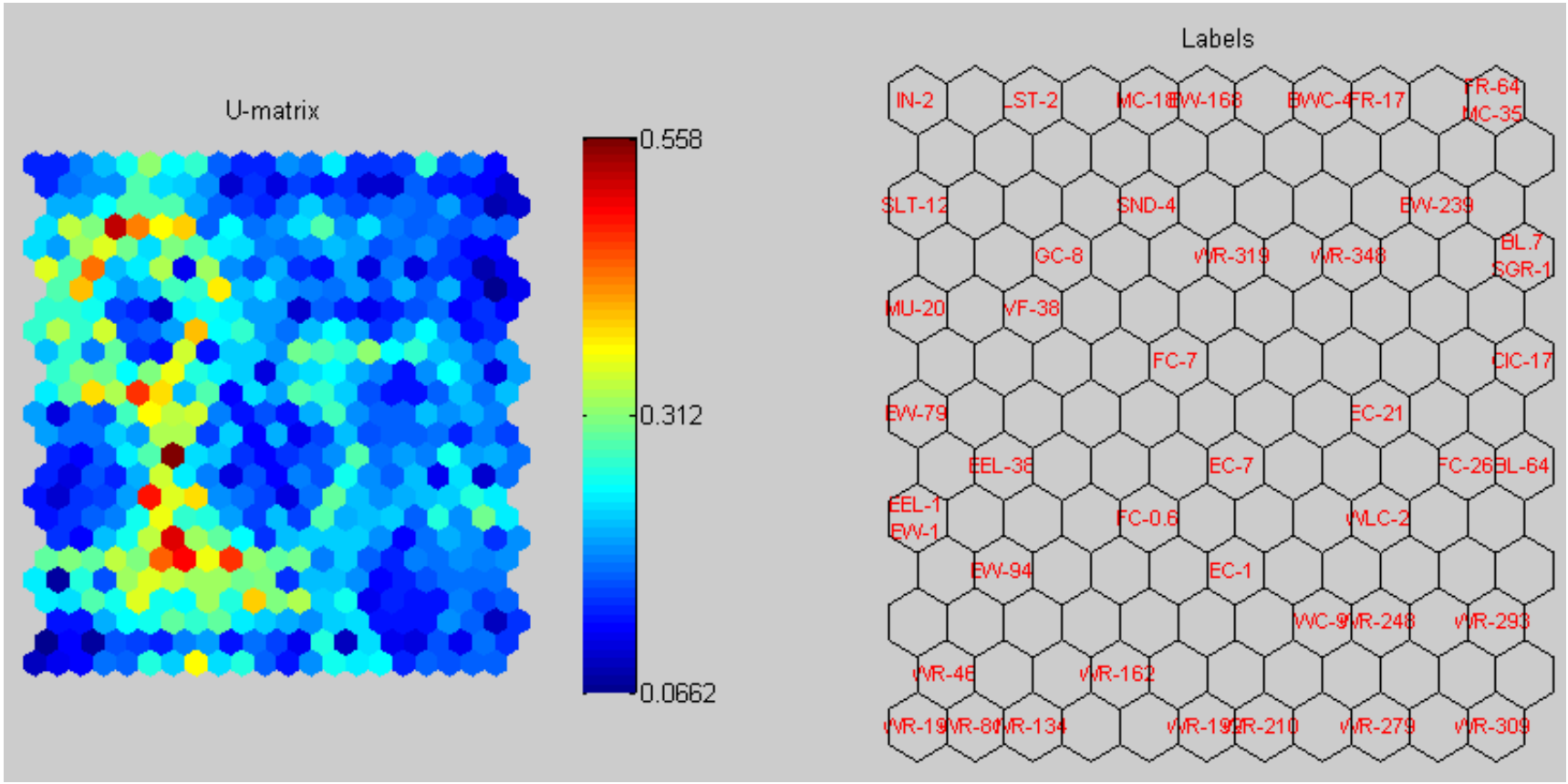
Supplementary Figure 3.7 – Quarter 1 Trimmed Mean dataset U-Matrix and station organization on the SOM



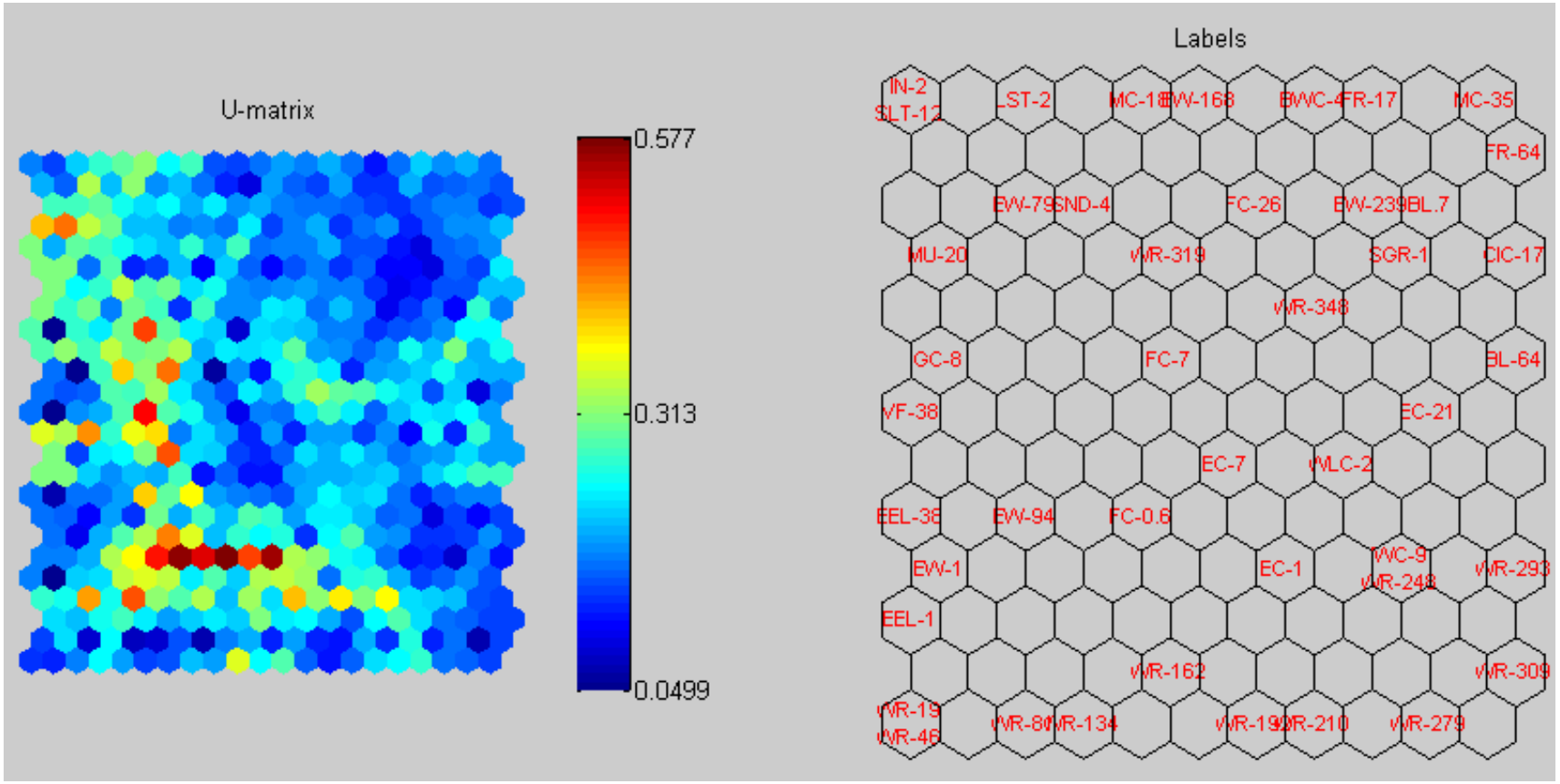
Supplementary Figure 3.8 – Quarter 1 Geometric Mean dataset U-Matrix and station organization on the SOM



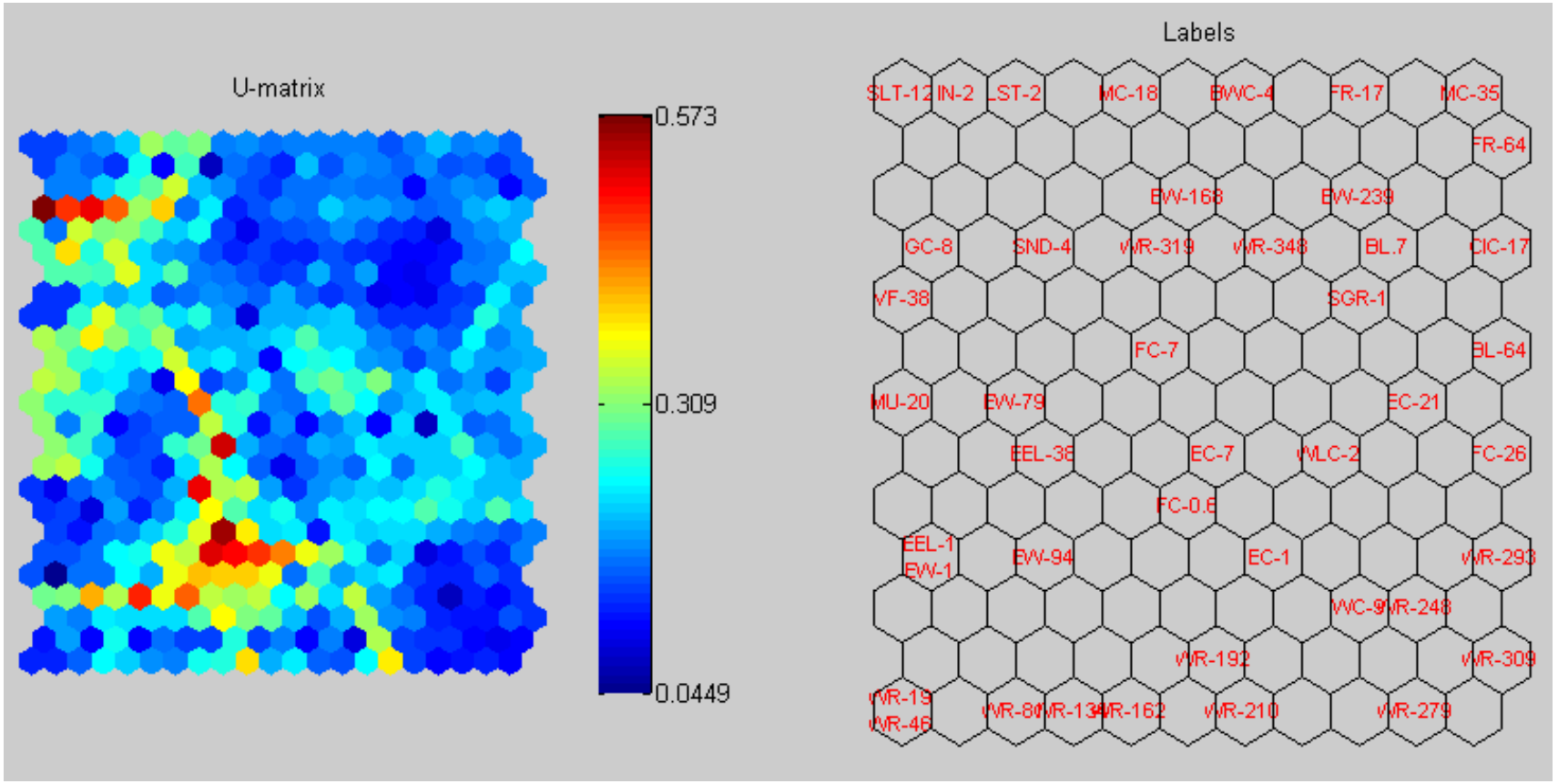
Supplementary Figure 3.9 – Quarter 2 Mean dataset U-Matrix and station organization on the SOM



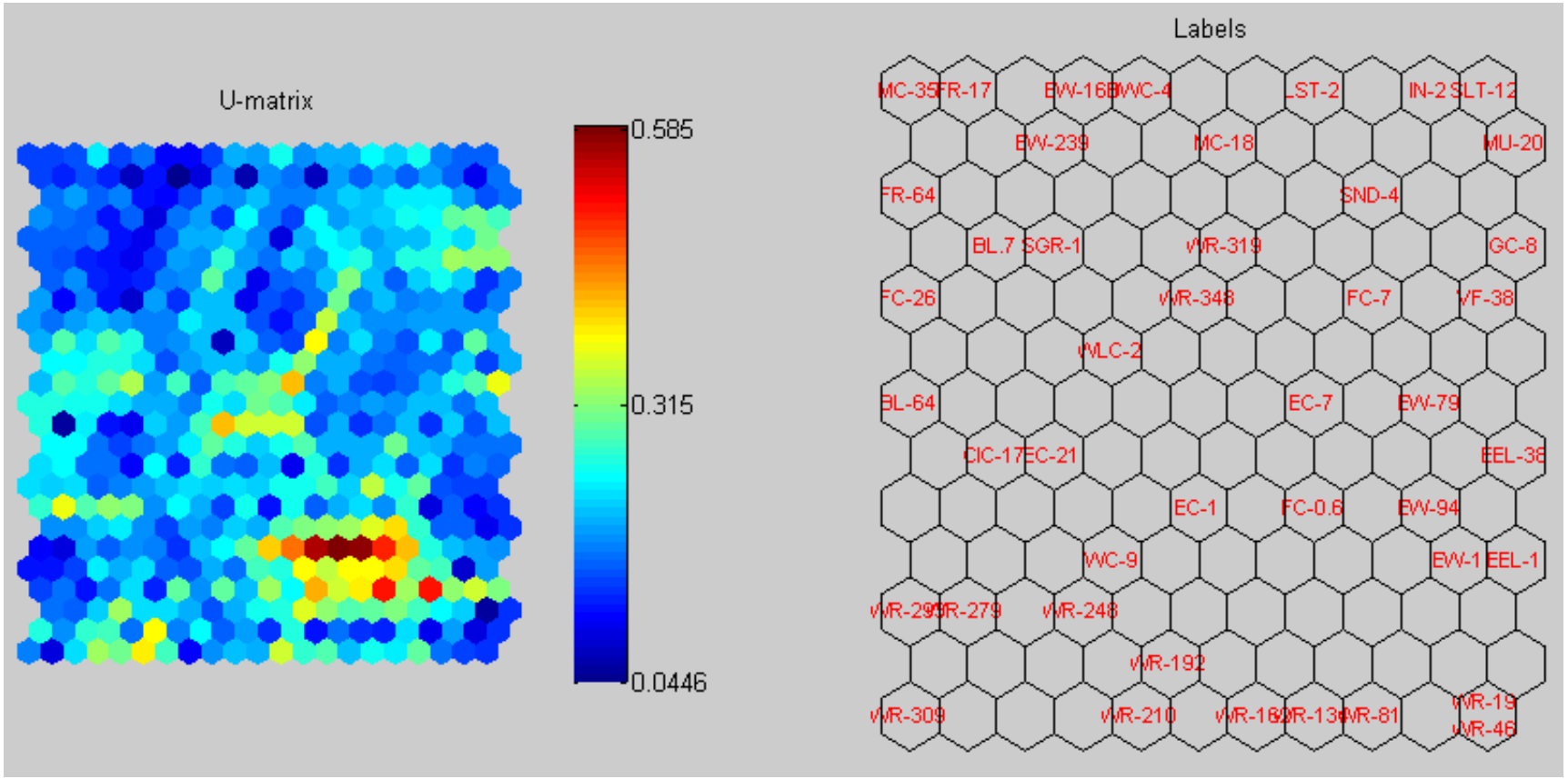
Supplementary Figure 3.10 – Quarter 2 Median dataset U-Matrix and station organization on the SOM



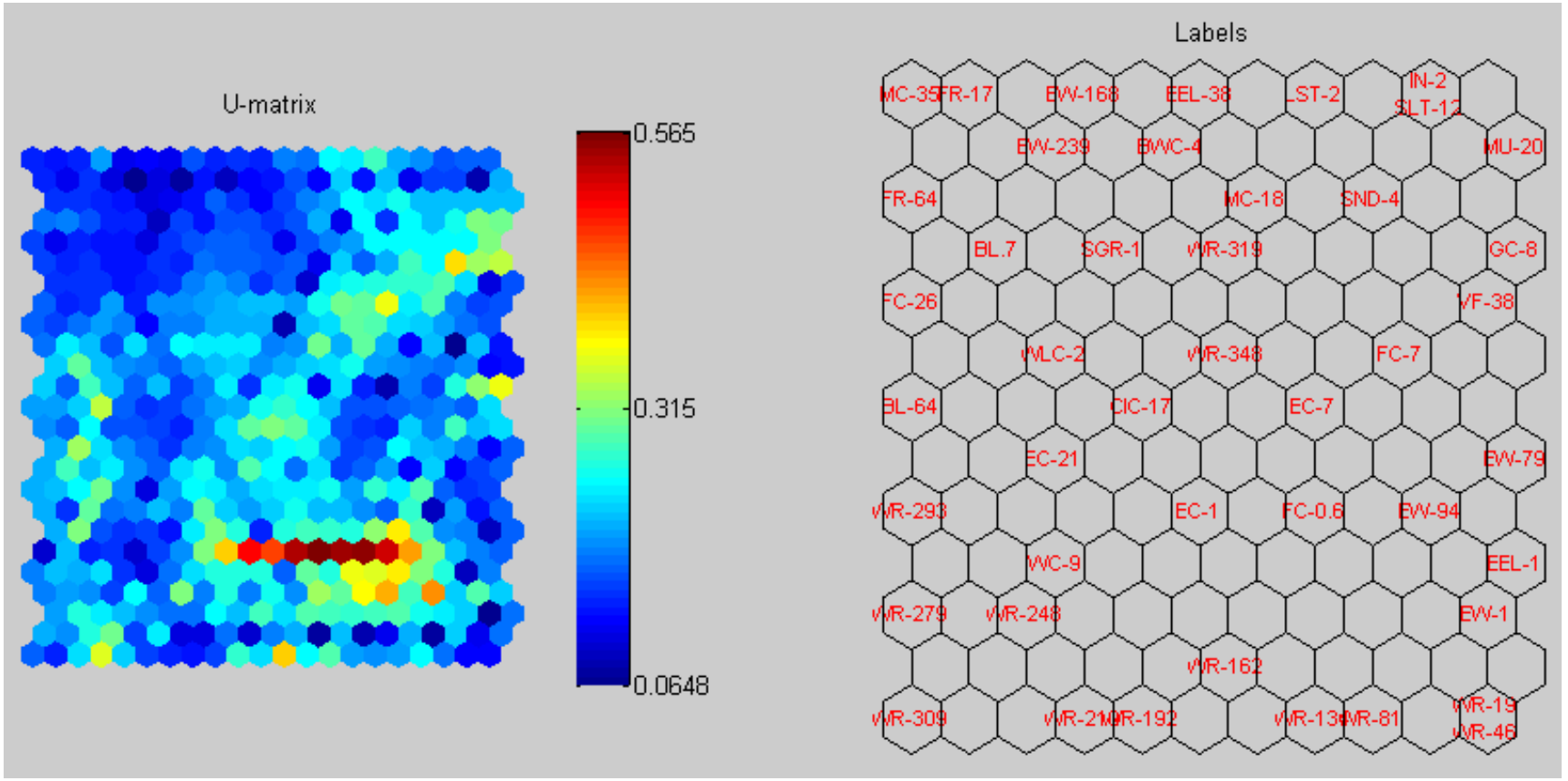
Supplementary Figure 3.11 – Quarter 2 Trimmed Mean dataset U-Matrix and station organization on the SOM



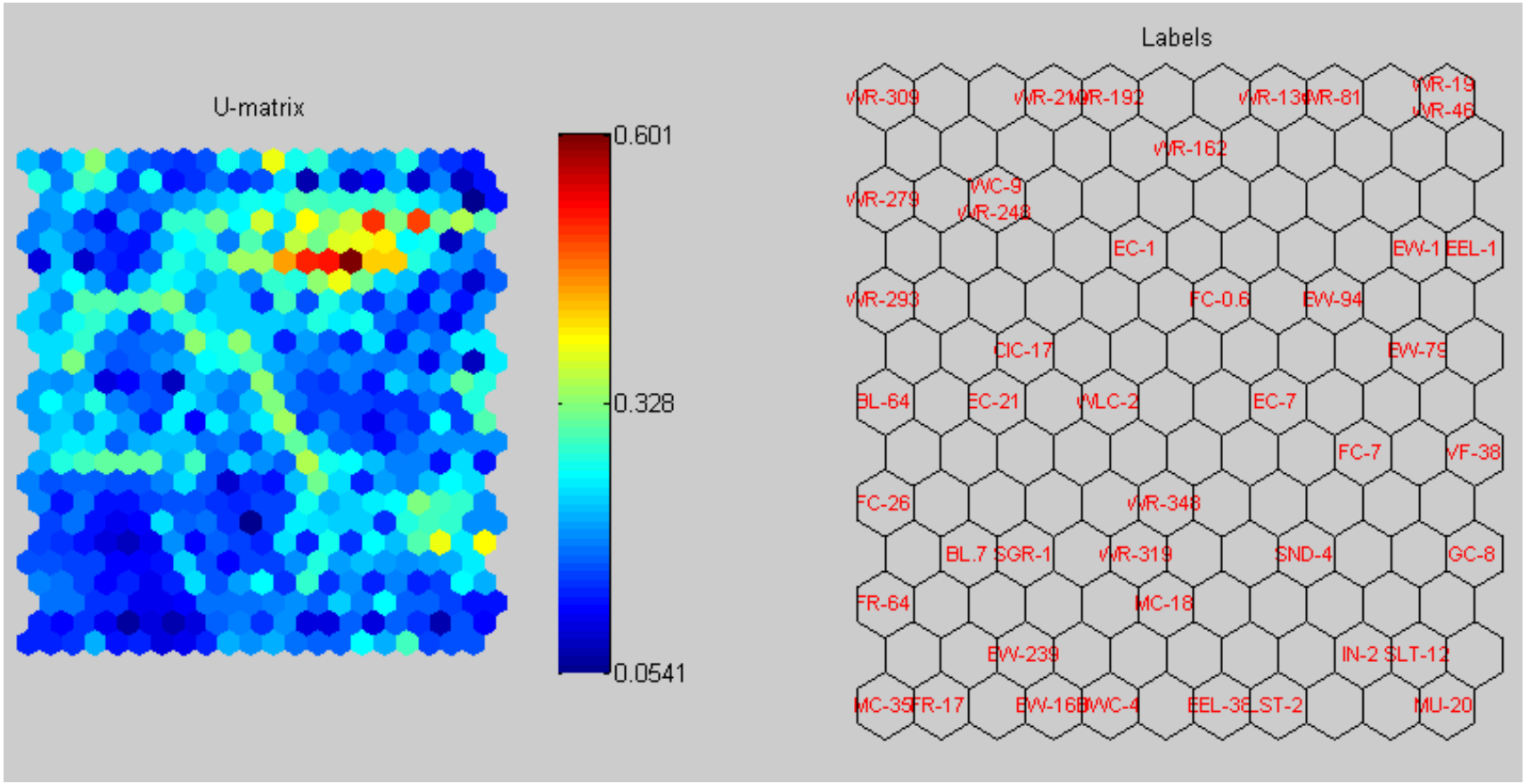
Supplementary Figure 3.12 – Quarter 2 Geometric Mean dataset U-Matrix and station organization on the SOM



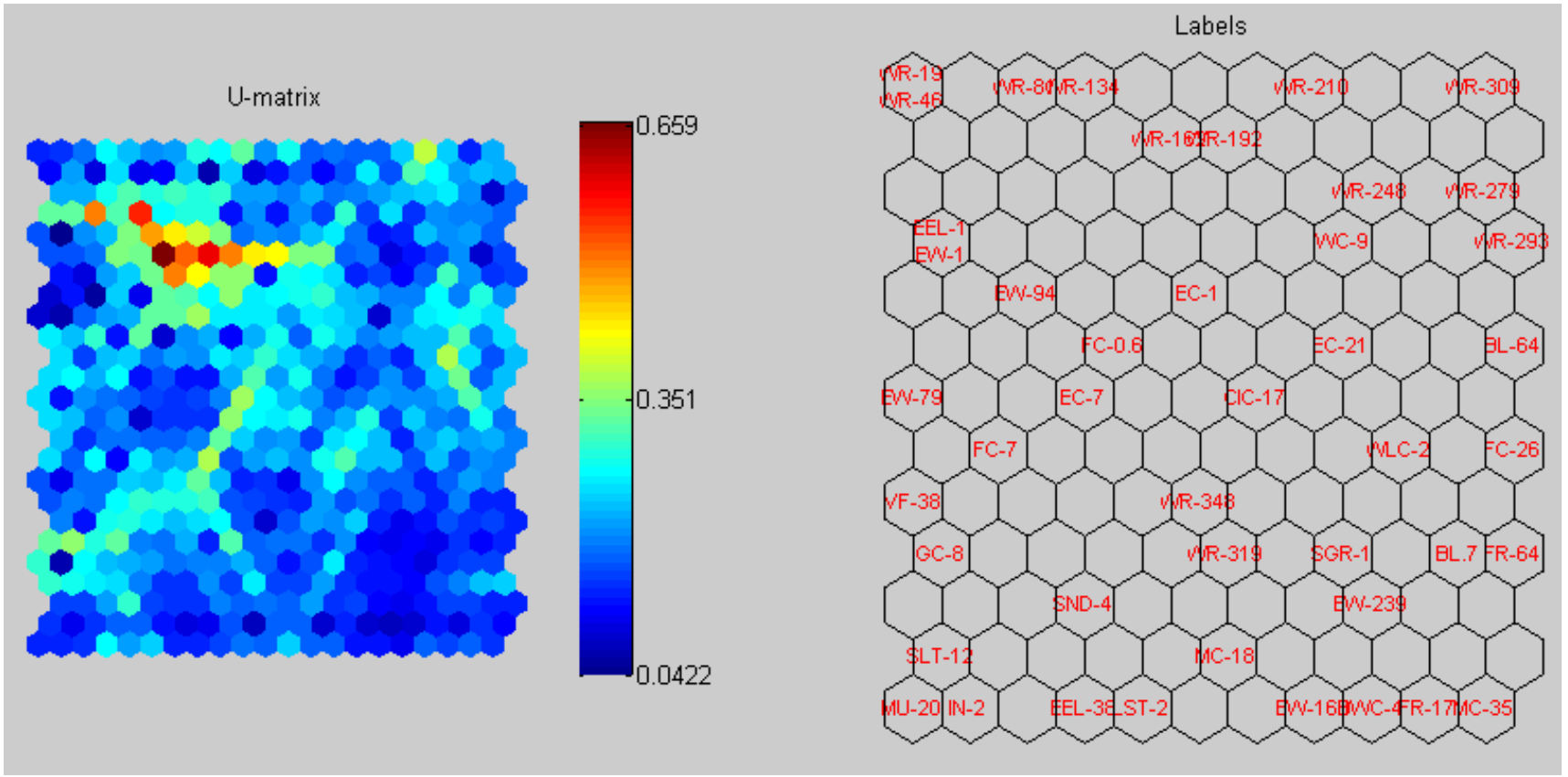
Supplementary Figure 3.13 – Quarter 3 Mean dataset U-Matrix and station organization on the SOM



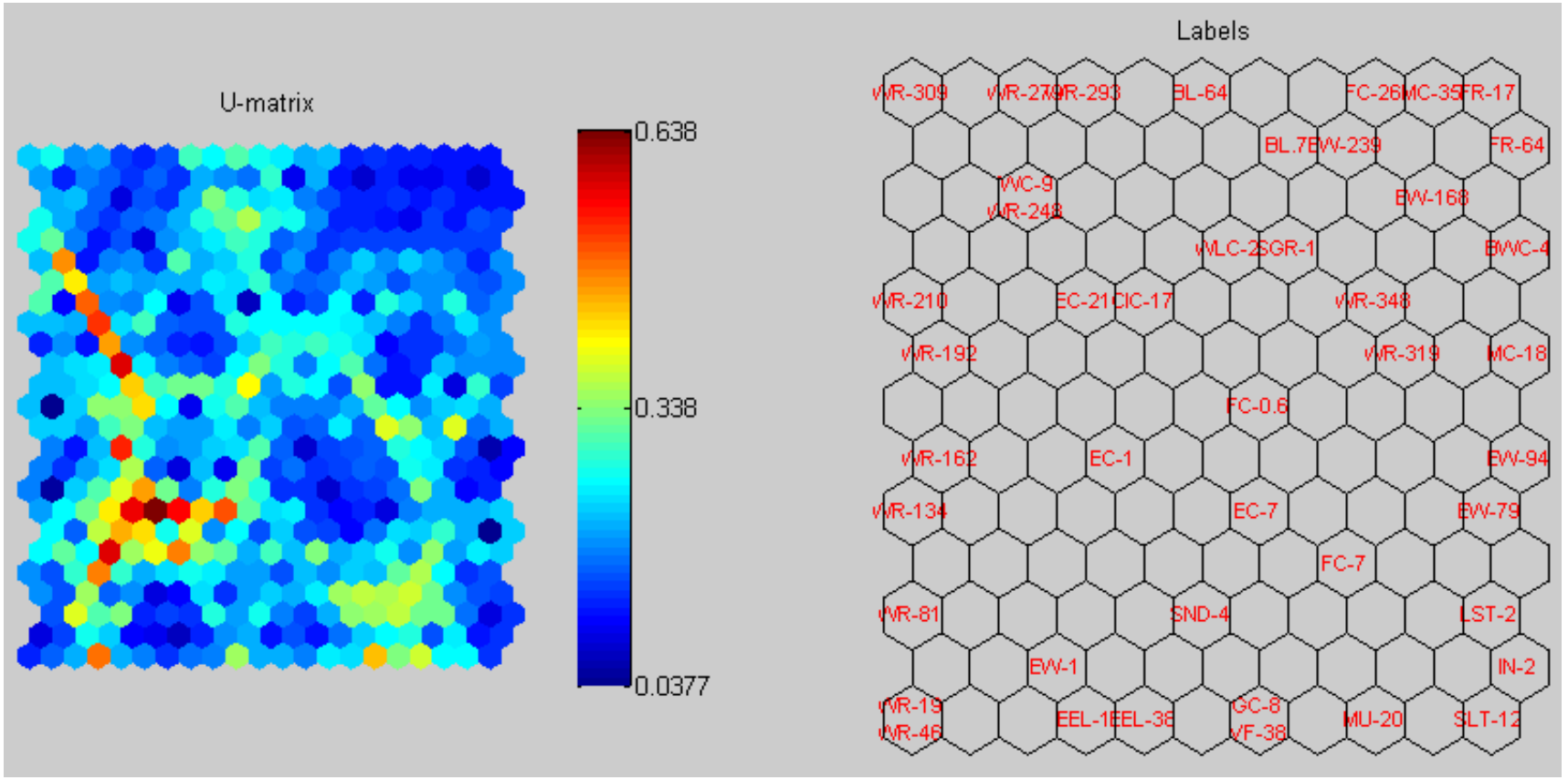
Supplementary Figure 3.14 – Quarter 3 Median dataset U-Matrix and station organization on the SOM



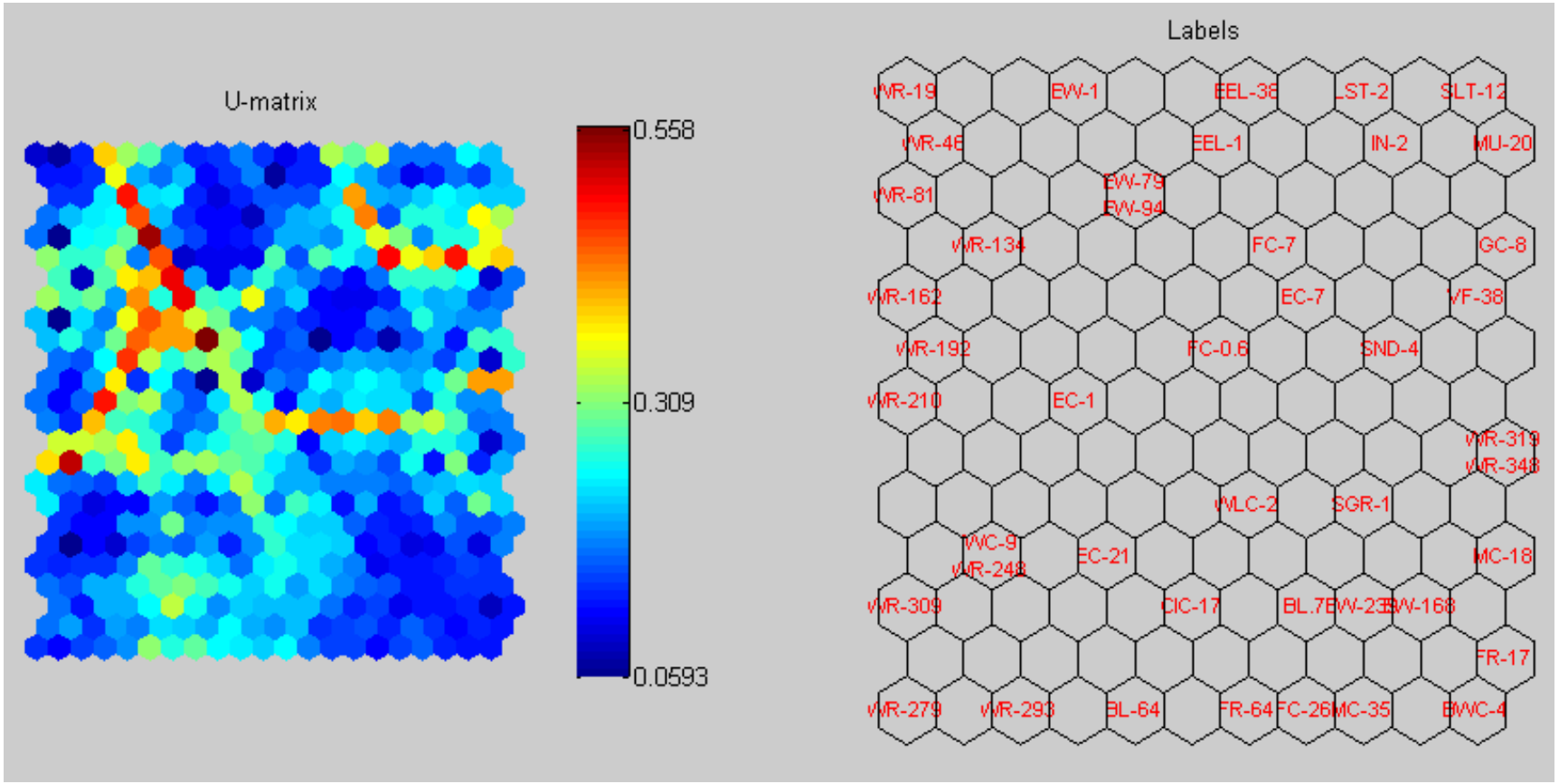
Supplementary Figure 3.15 – Quarter 3 Trimmed Mean dataset U-Matrix and station organization on the SOM



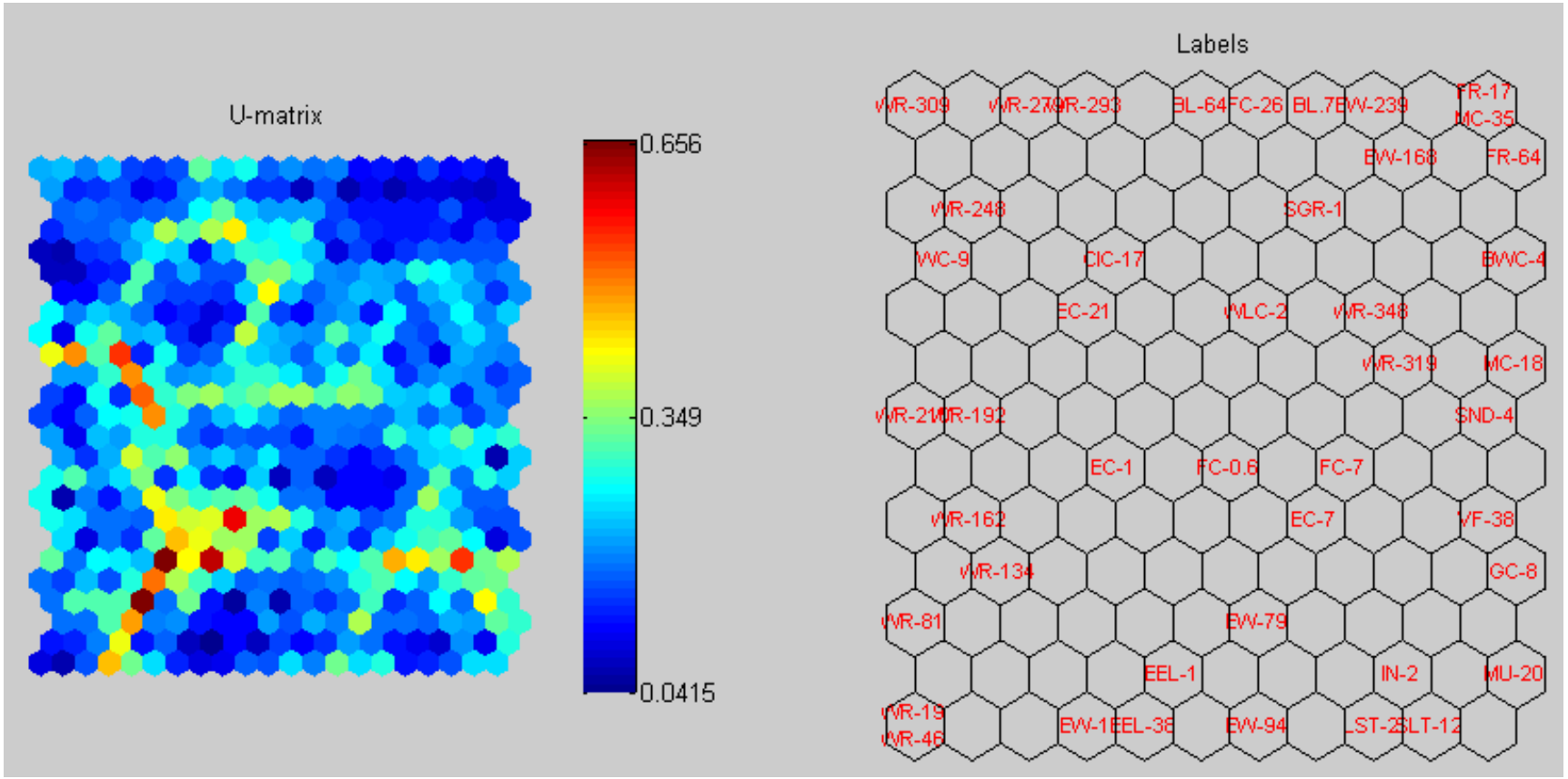
Supplementary Figure 3.16 – Quarter 3 Geometric Mean dataset U-Matrix and station organization on the SOM



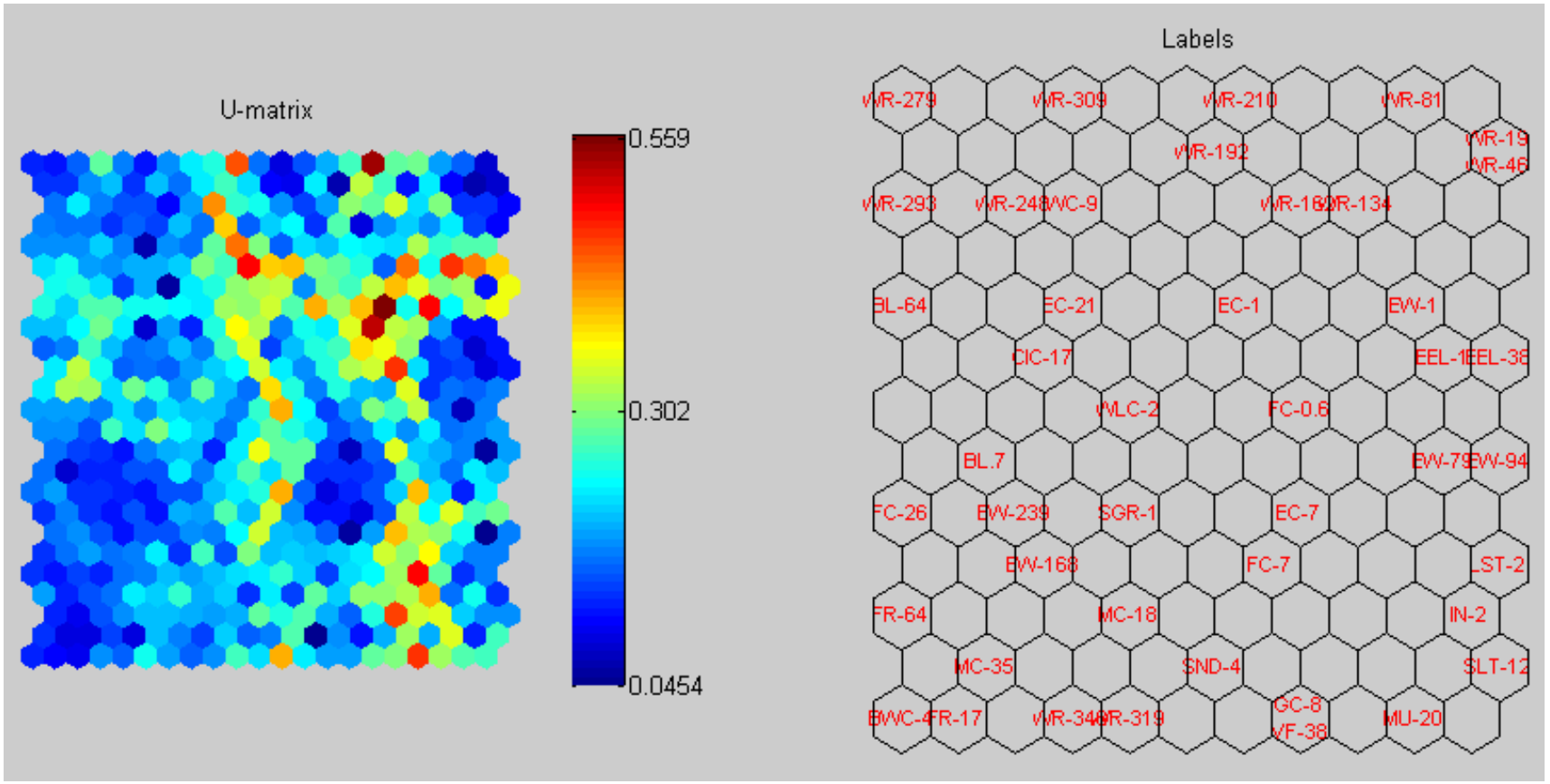
Supplementary Figure 3.17 – Quarter 4 Mean dataset U-Matrix and station organization on the SOM



Supplementary Figure 3.18 – Quarter 4 Median dataset U-Matrix and station organization on the SOM

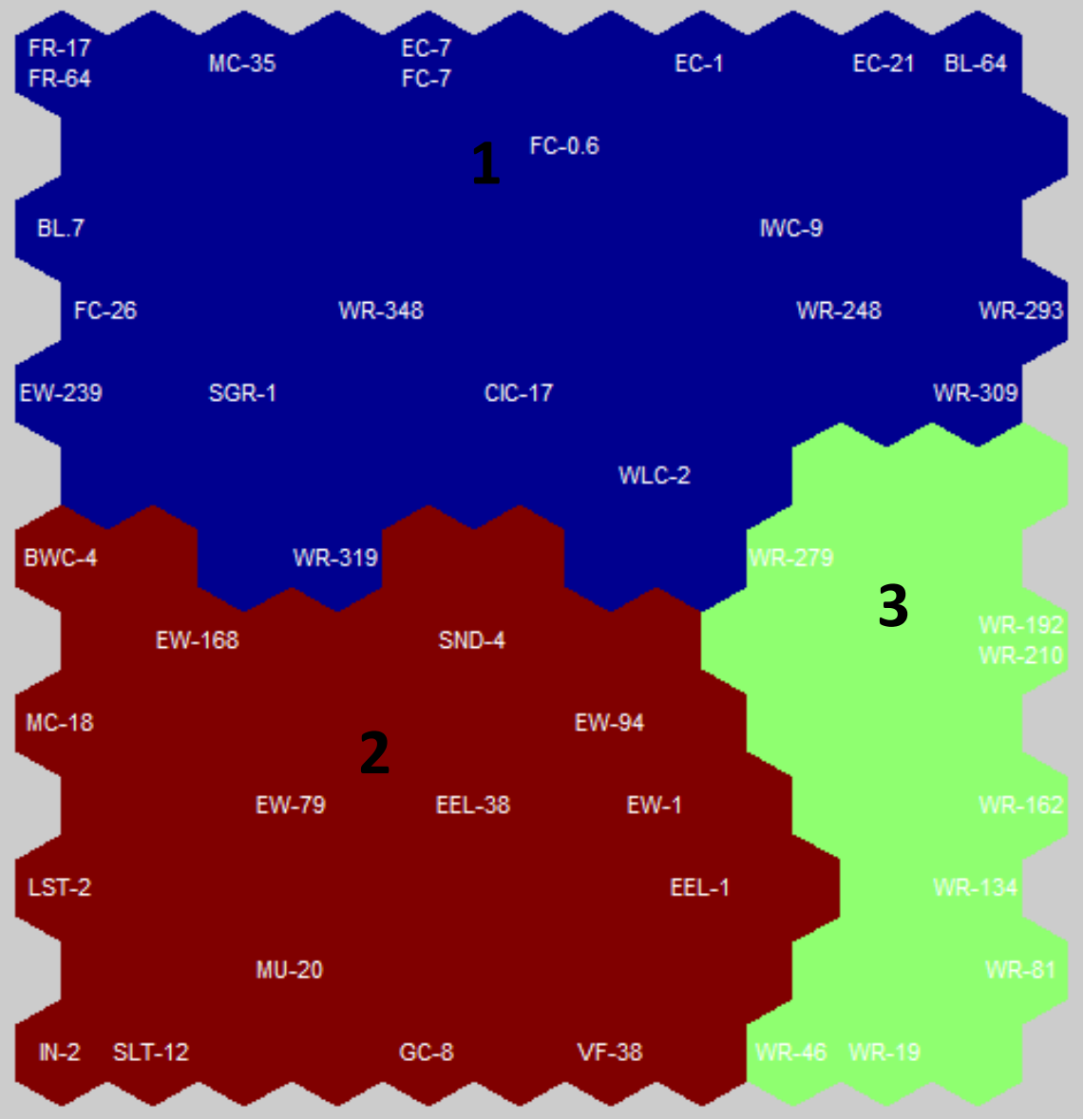


Supplementary Figure 3.19 – Quarter 4 Trimmed Mean dataset U-Matrix and station organization on the SOM

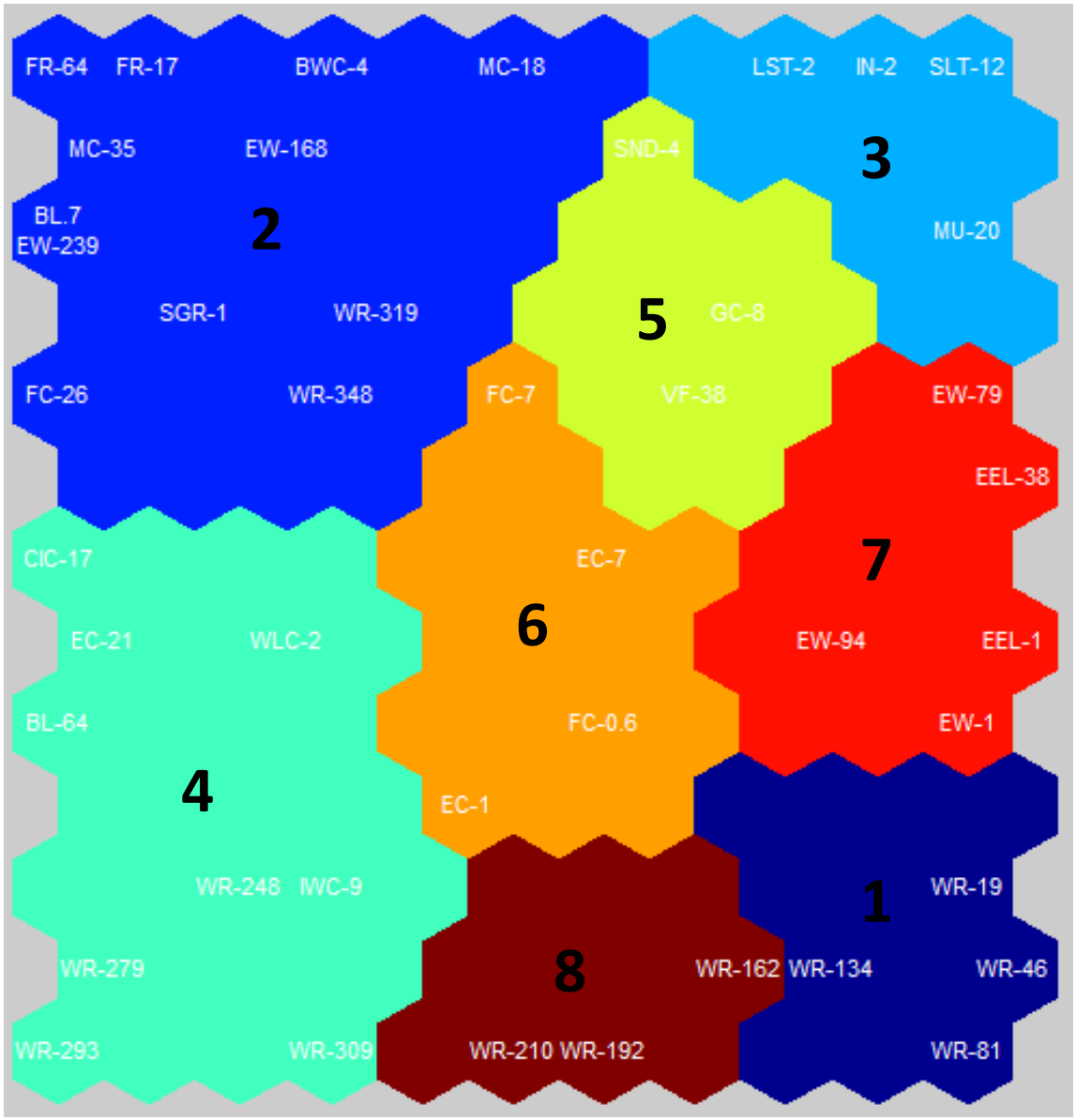


Supplementary Figure 3.20 – Quarter 4 Geometric Mean dataset U-Matrix and station organization on the SOM

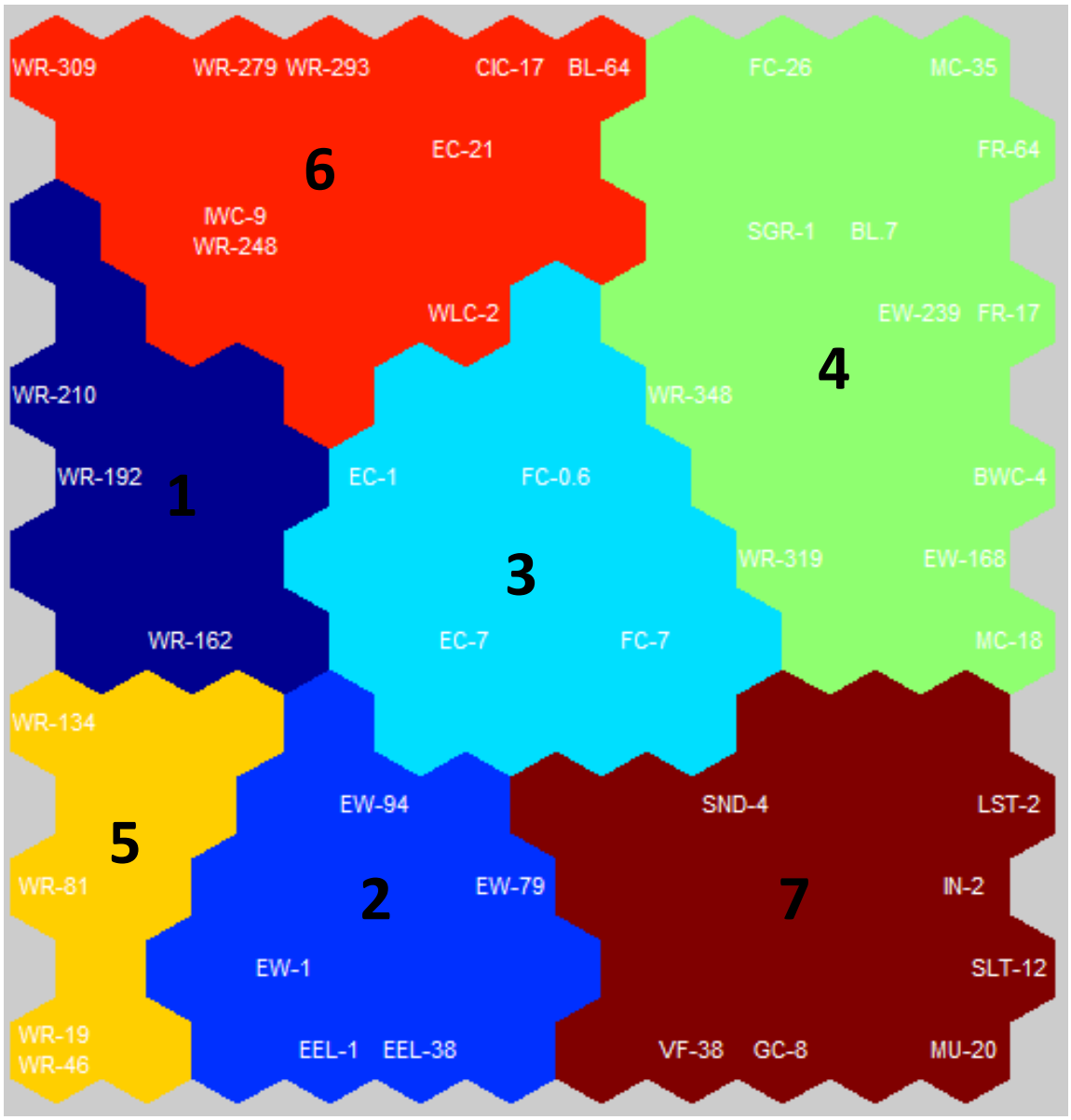
SOM Cluster Arrangements



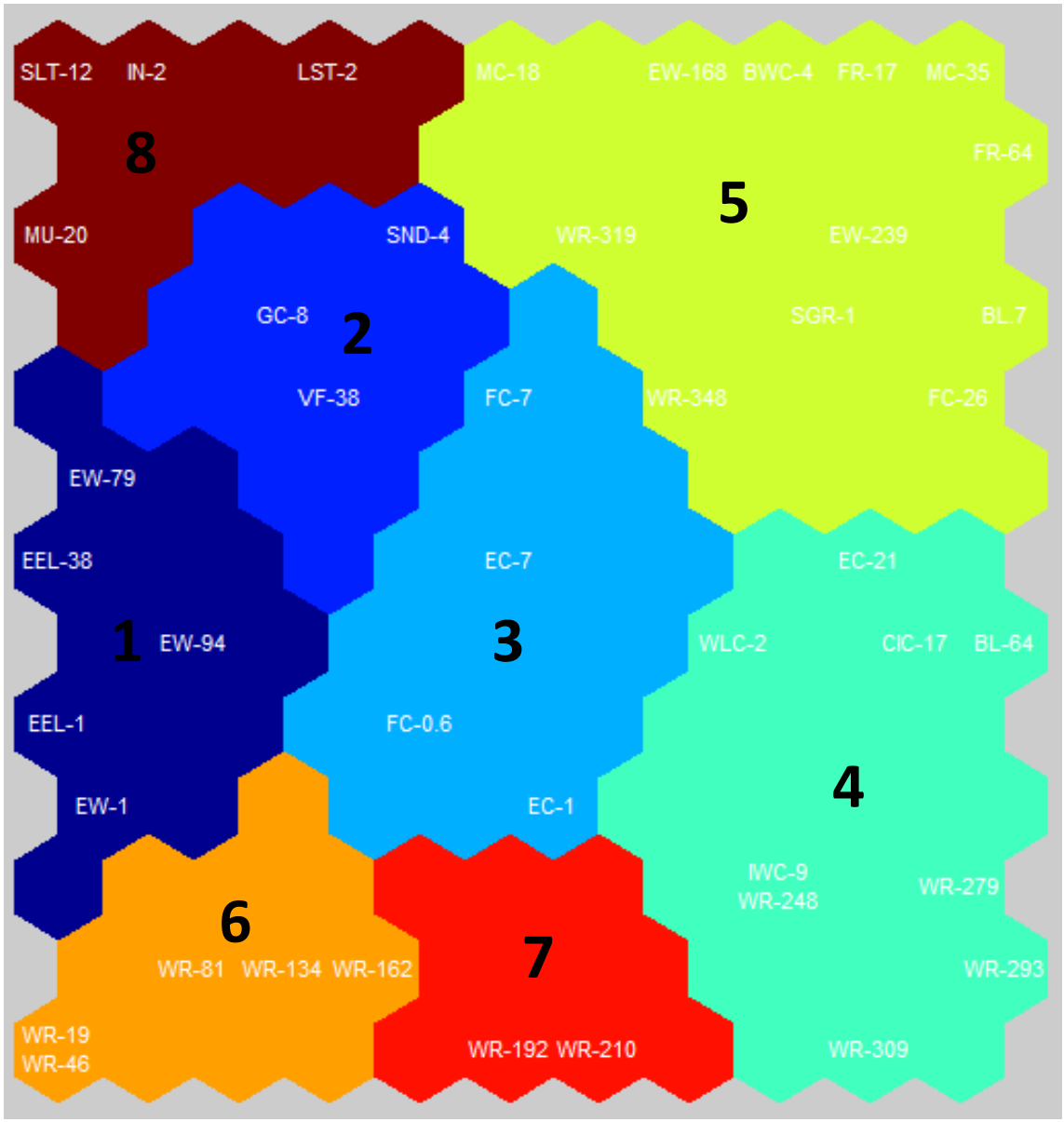
Supplementary Figure 4.1 – SOM cluster configuration for the annual mean dataset



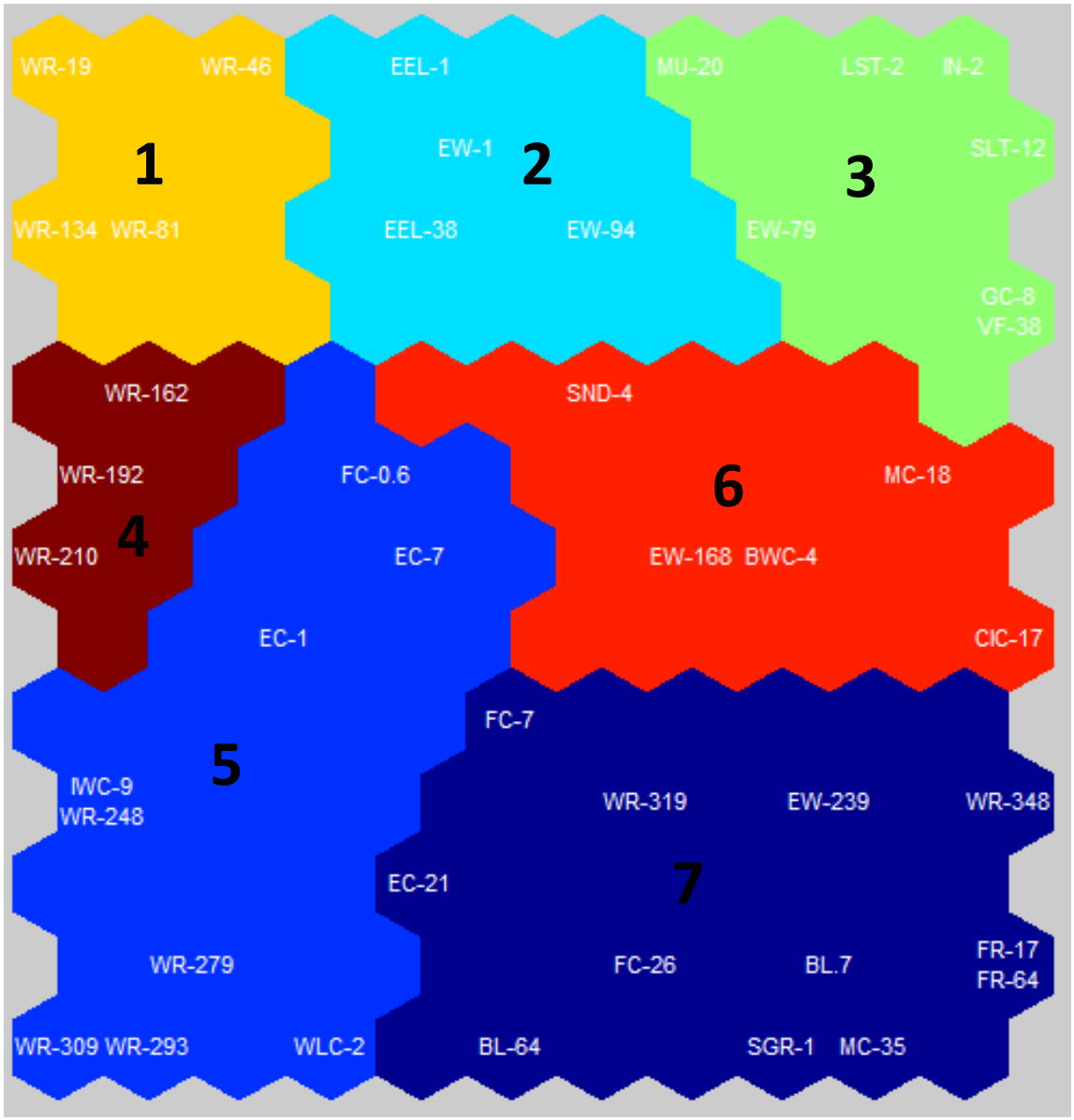
Supplementary Figure 4.2 – SOM cluster configuration for the annual median dataset



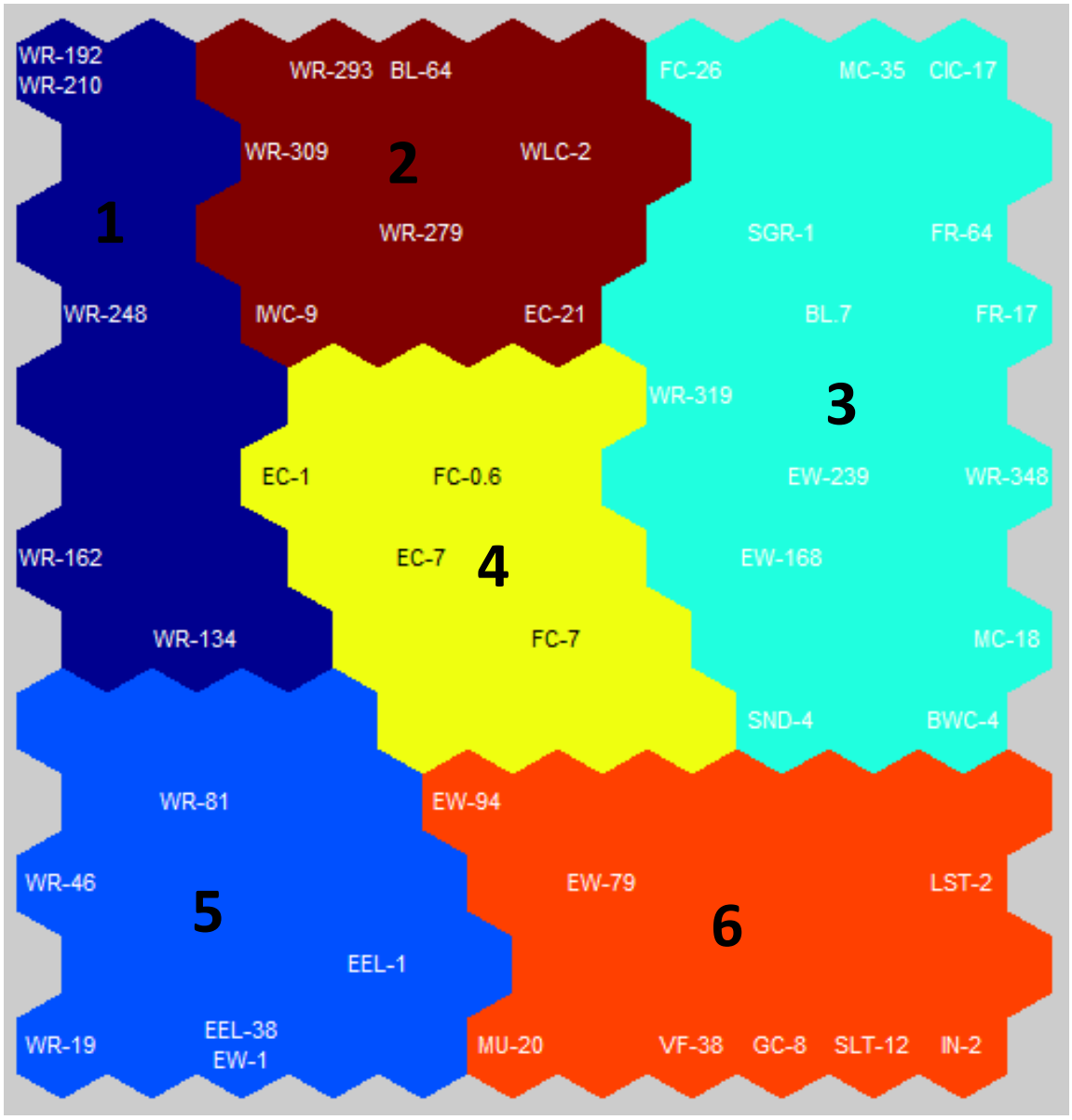
Supplementary Figure 4.3 – SOM cluster configuration for the annual trimmed mean dataset



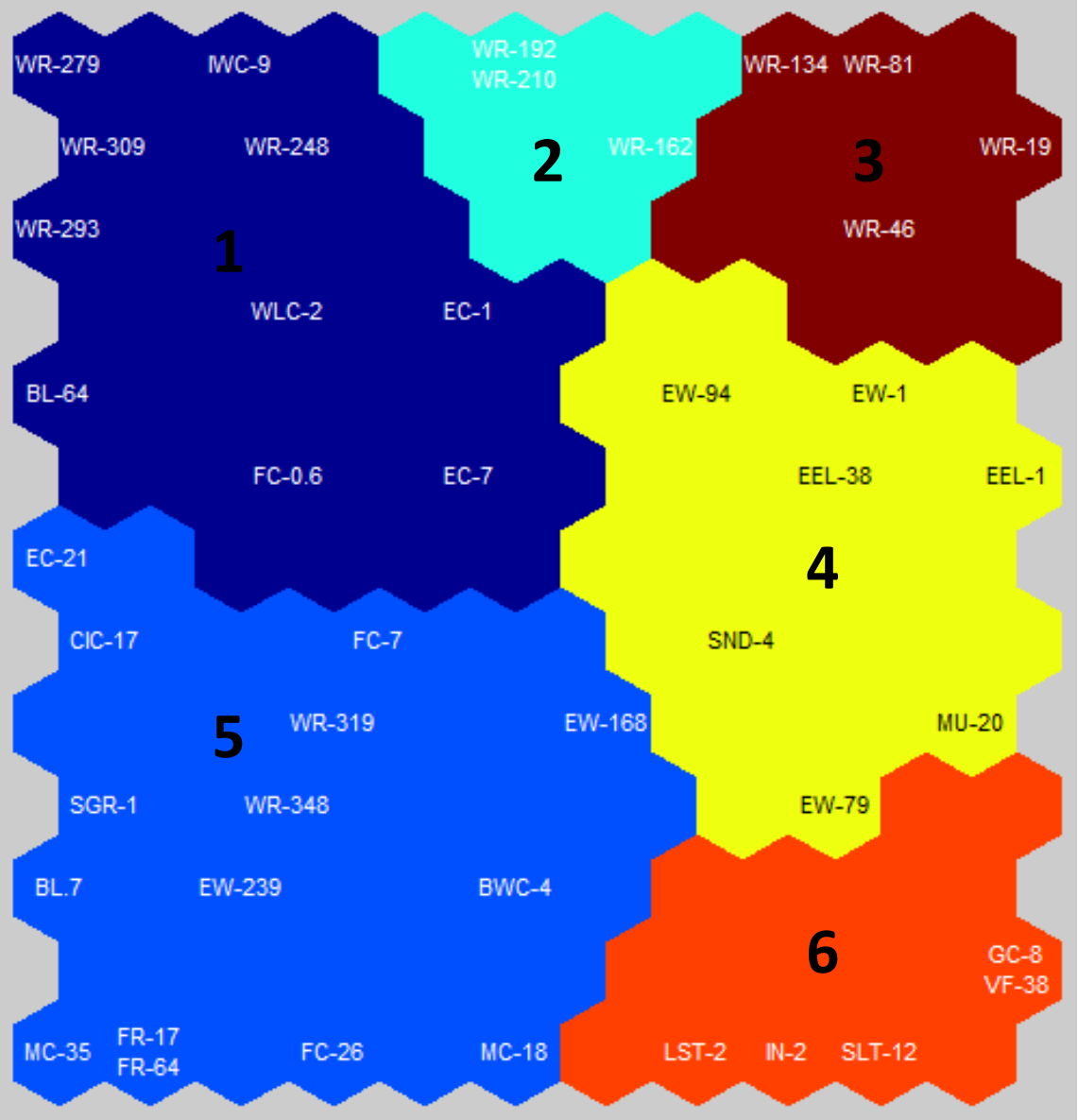
Supplementary Figure 4.4 – SOM cluster configuration for the annual geometric mean dataset



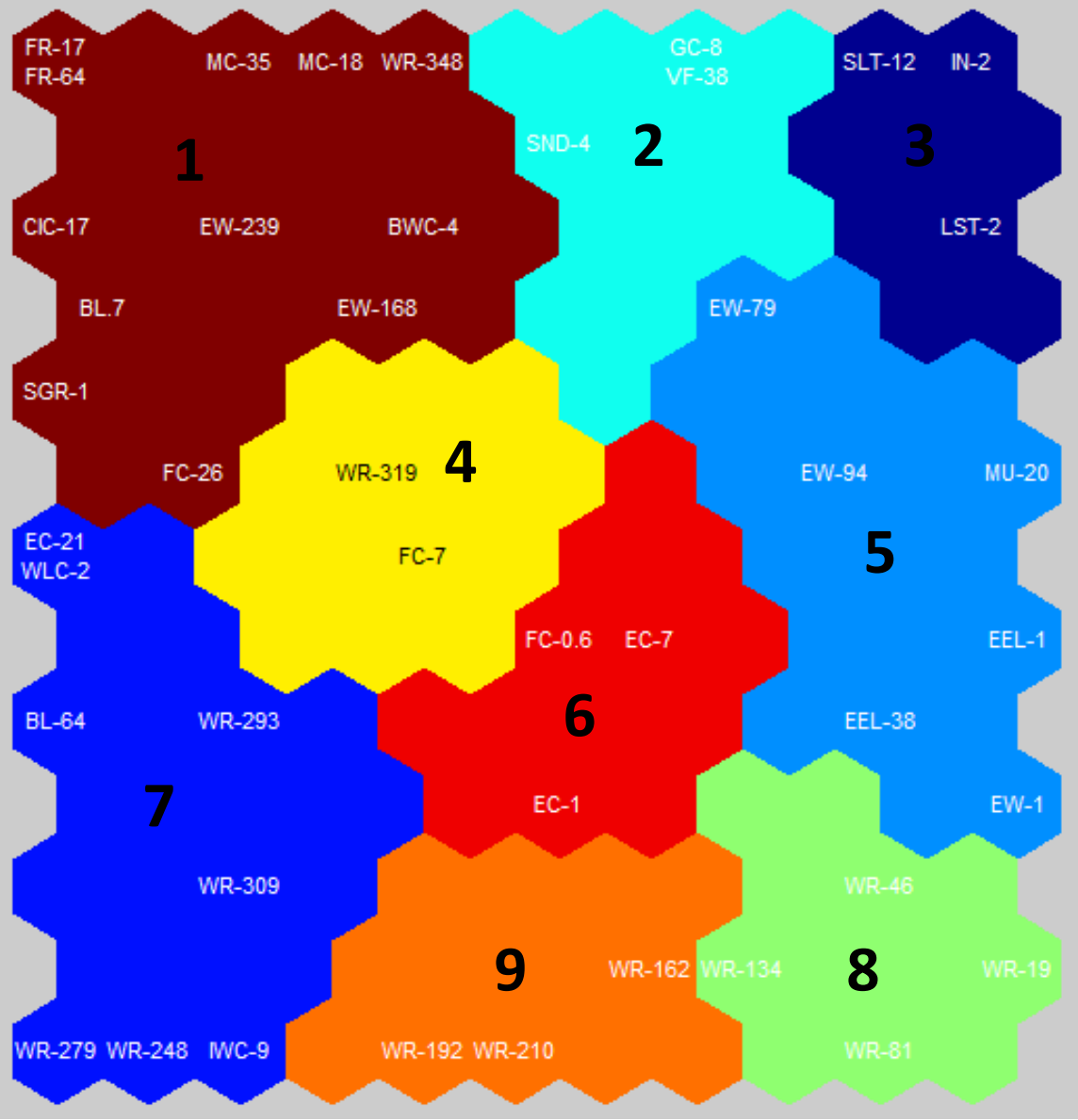
Supplementary Figure 4.5 – SOM cluster configuration for the quarter 1 mean dataset



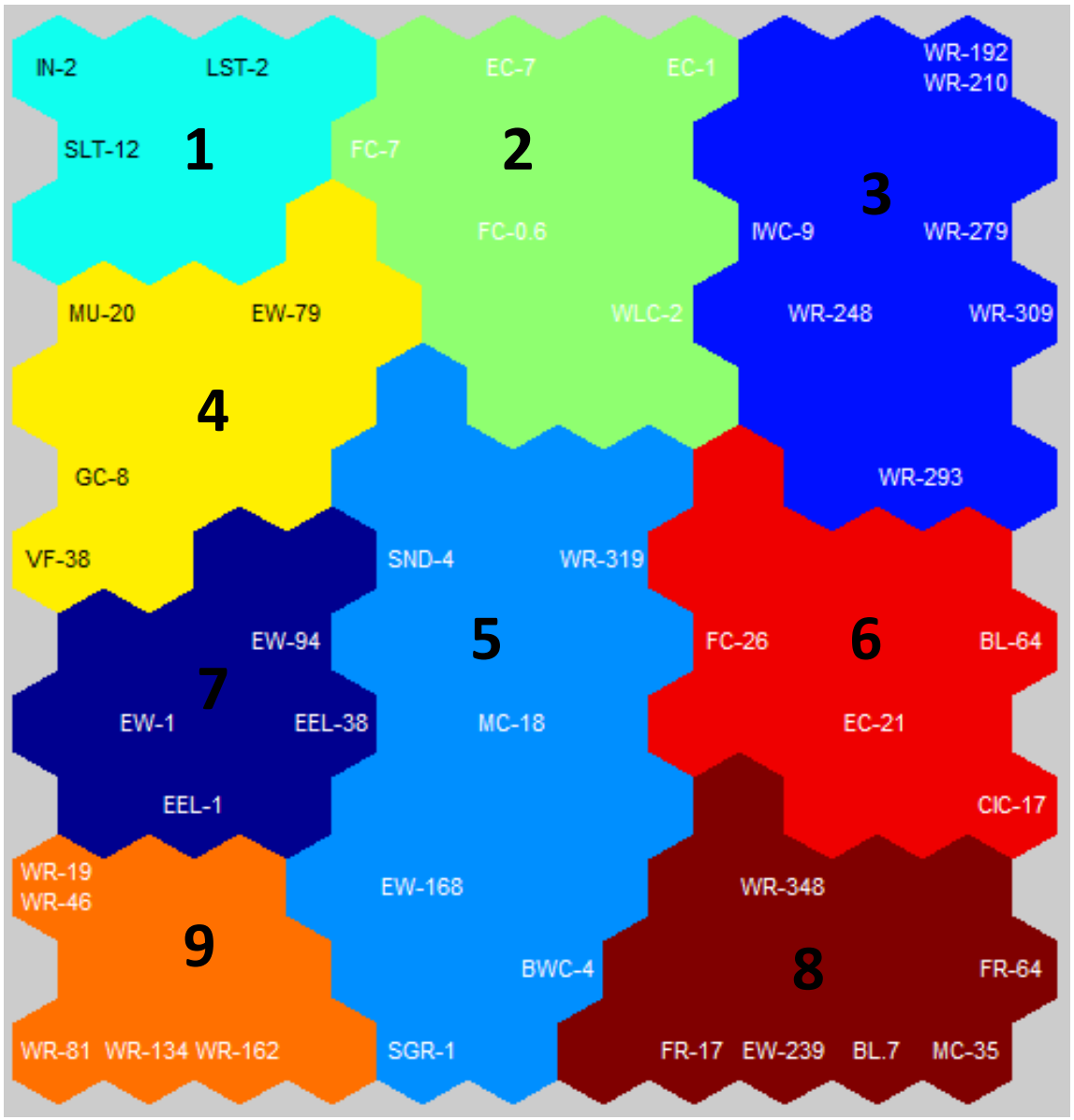
Supplementary Figure 4.6 – SOM cluster configuration for the quarter 1 median dataset



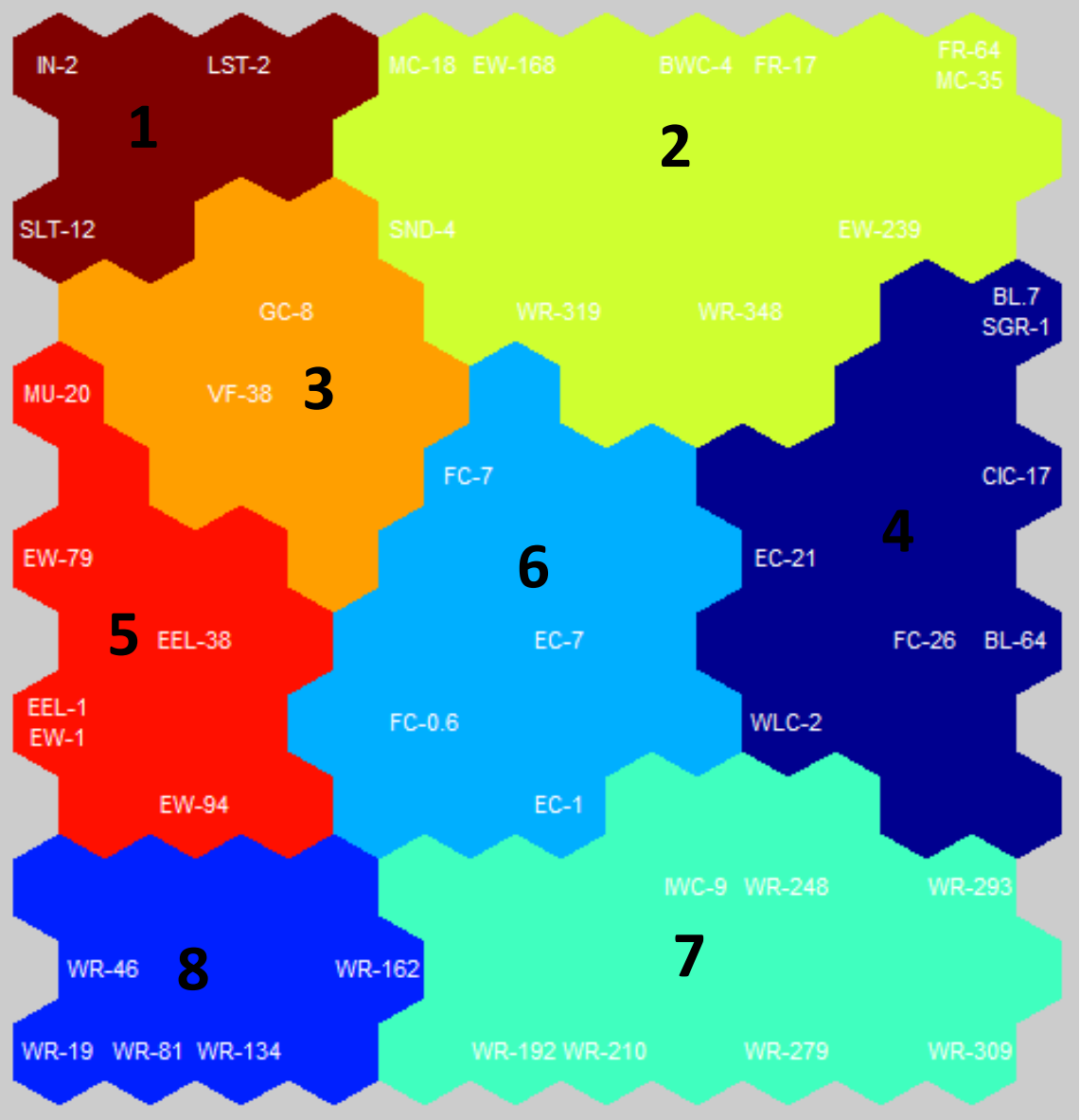
Supplementary Figure 4.7 – SOM cluster configuration for the quarter 1 trimmed mean dataset



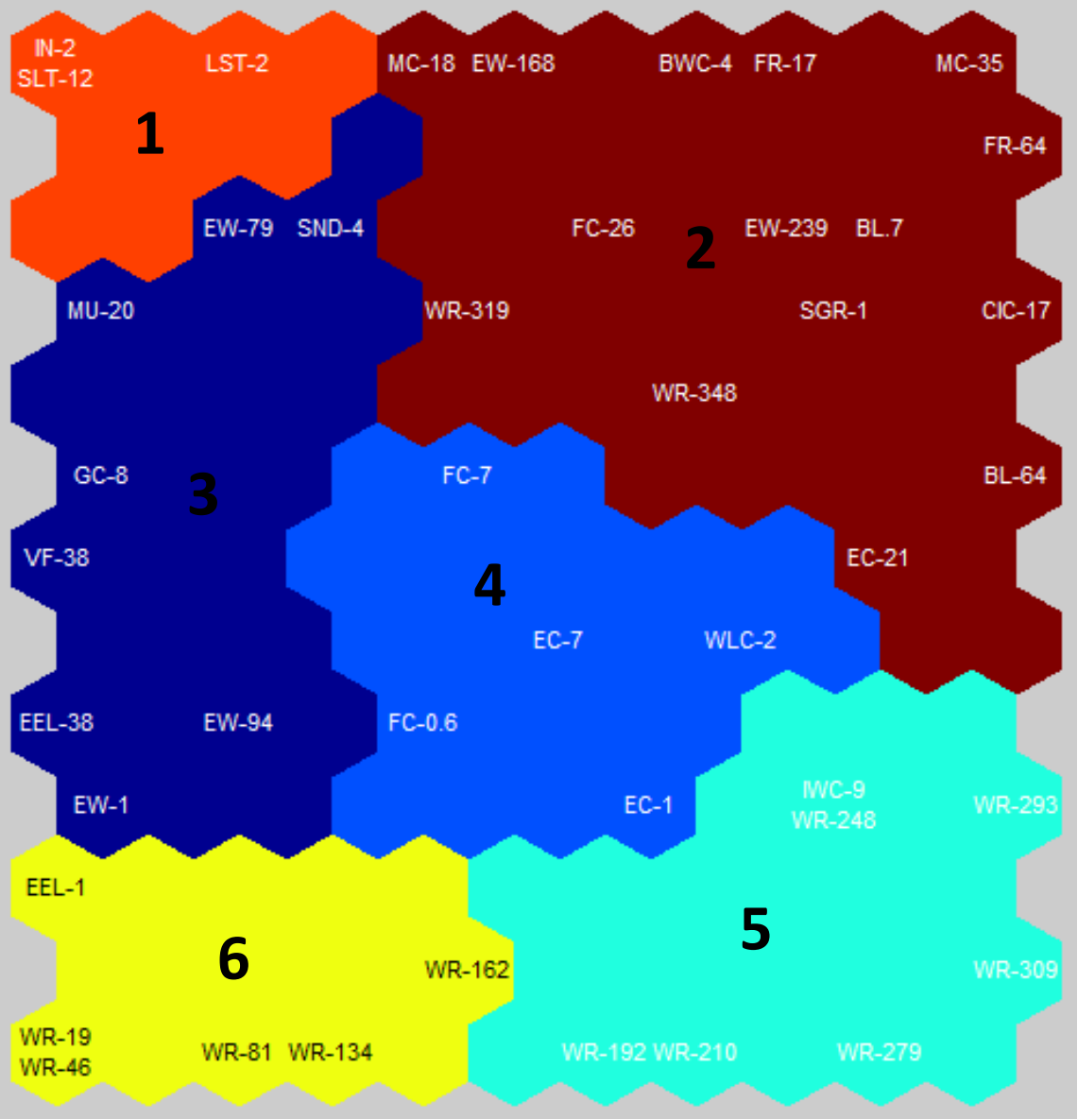
Supplementary Figure 4.8 – SOM cluster configuration for the quarter 1 geometric mean dataset



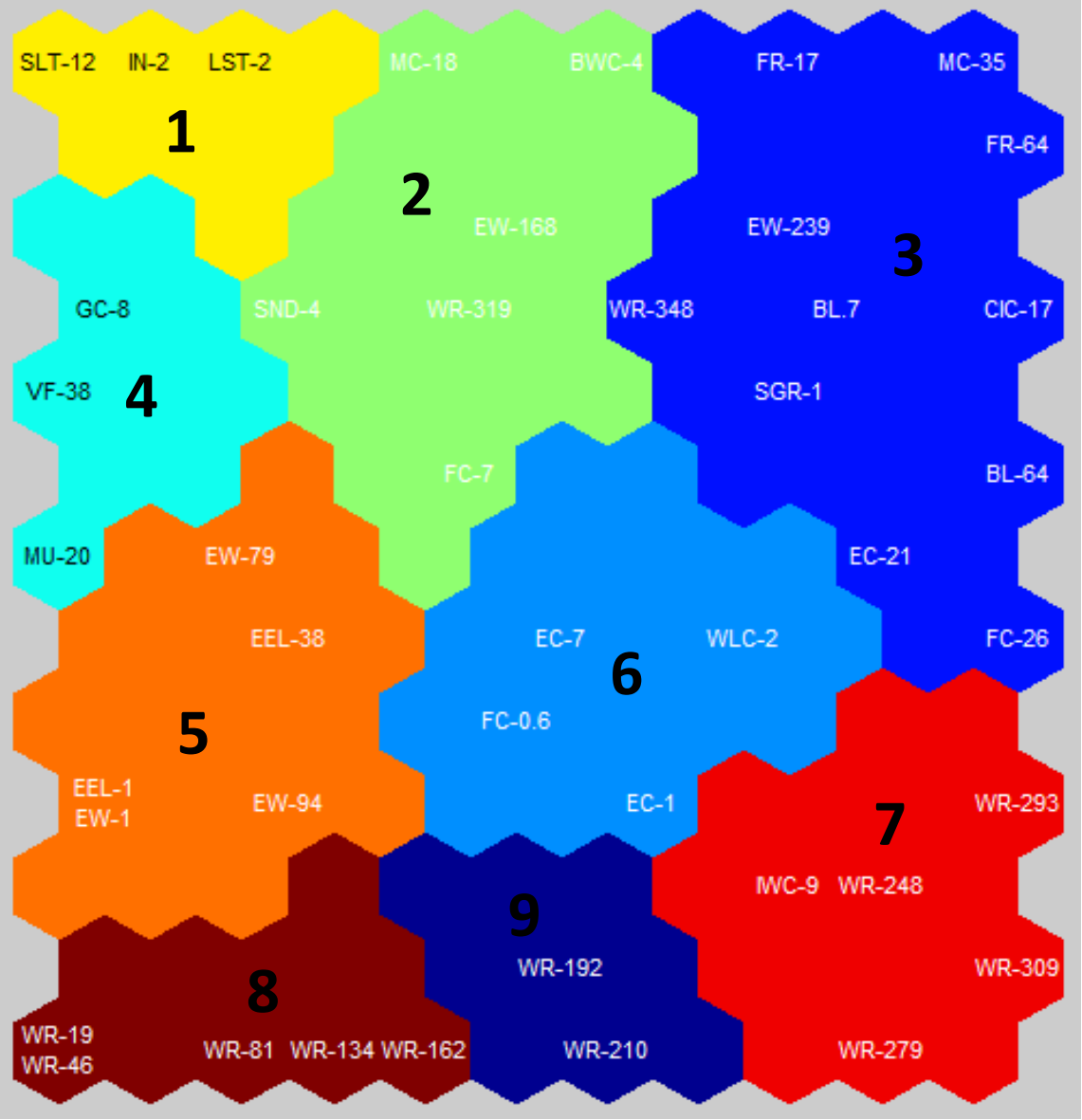
Supplementary Figure 4.9 – SOM cluster configuration for the quarter 2 mean dataset



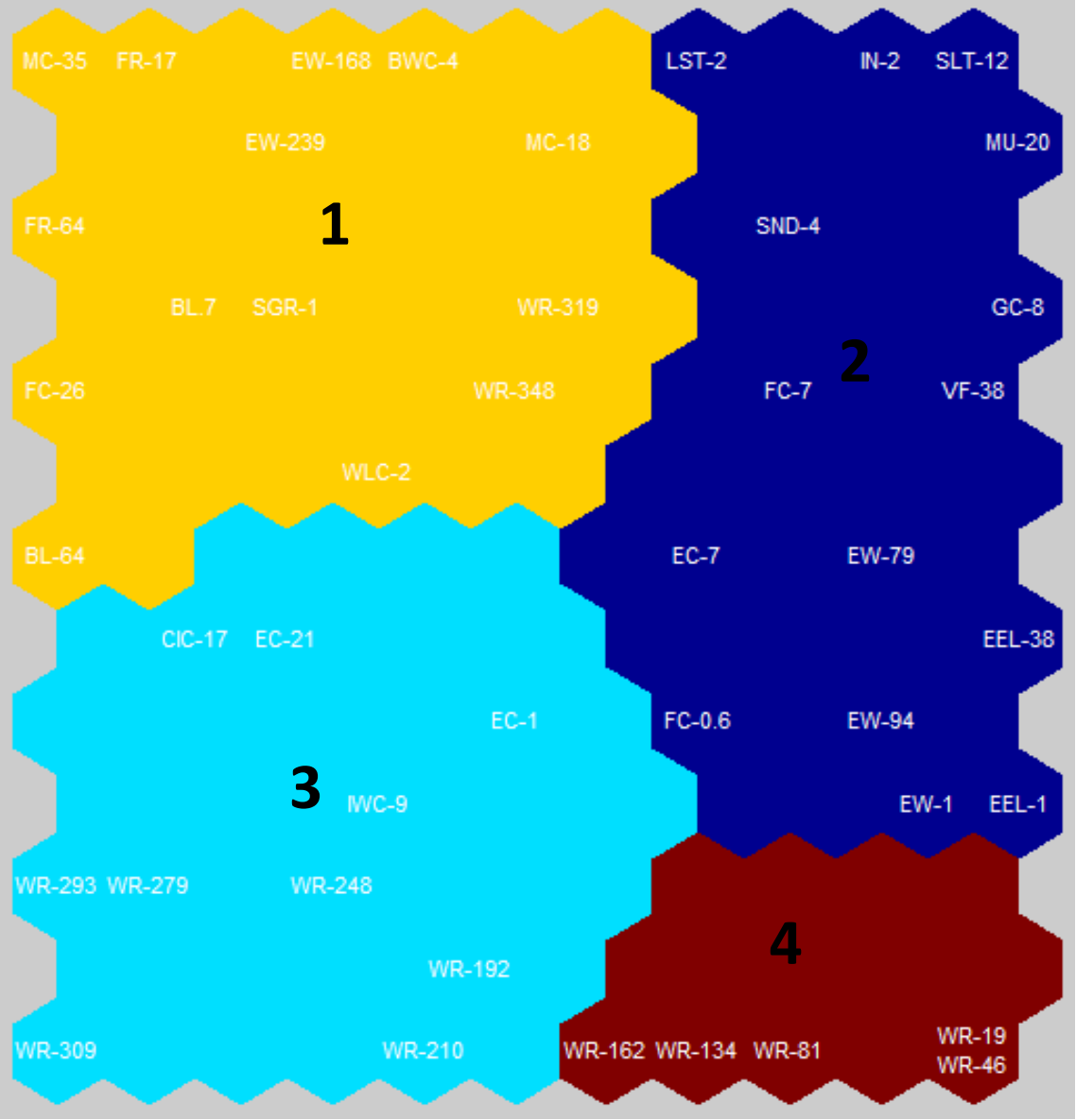
Supplementary Figure 4.10 – SOM cluster configuration for the quarter 2 median dataset



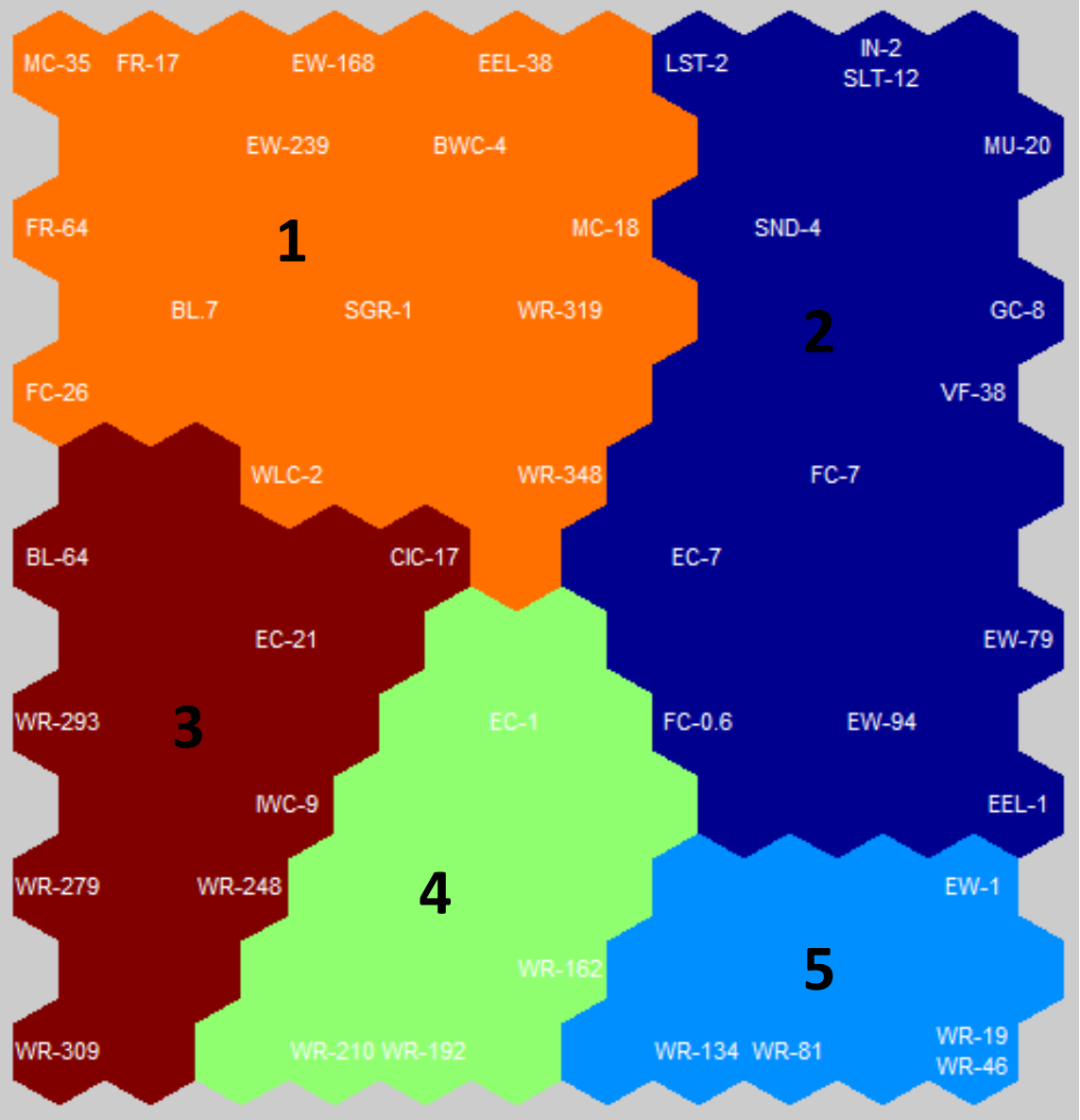
Supplementary Figure 4.11 – SOM cluster configuration for the quarter 2 trimmed mean dataset



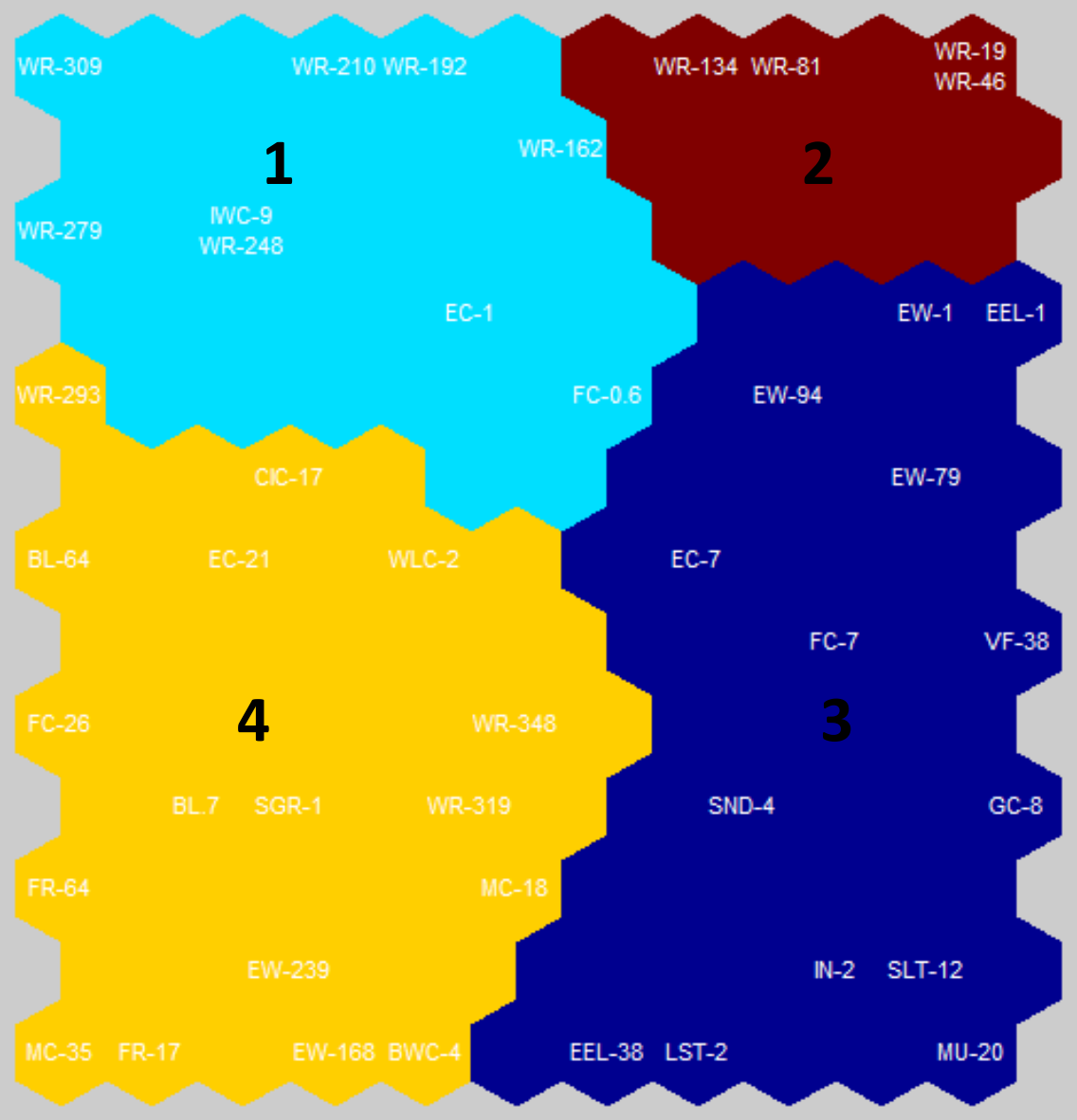
Supplementary Figure 4.12 – SOM cluster configuration for the quarter 2 geometric mean dataset



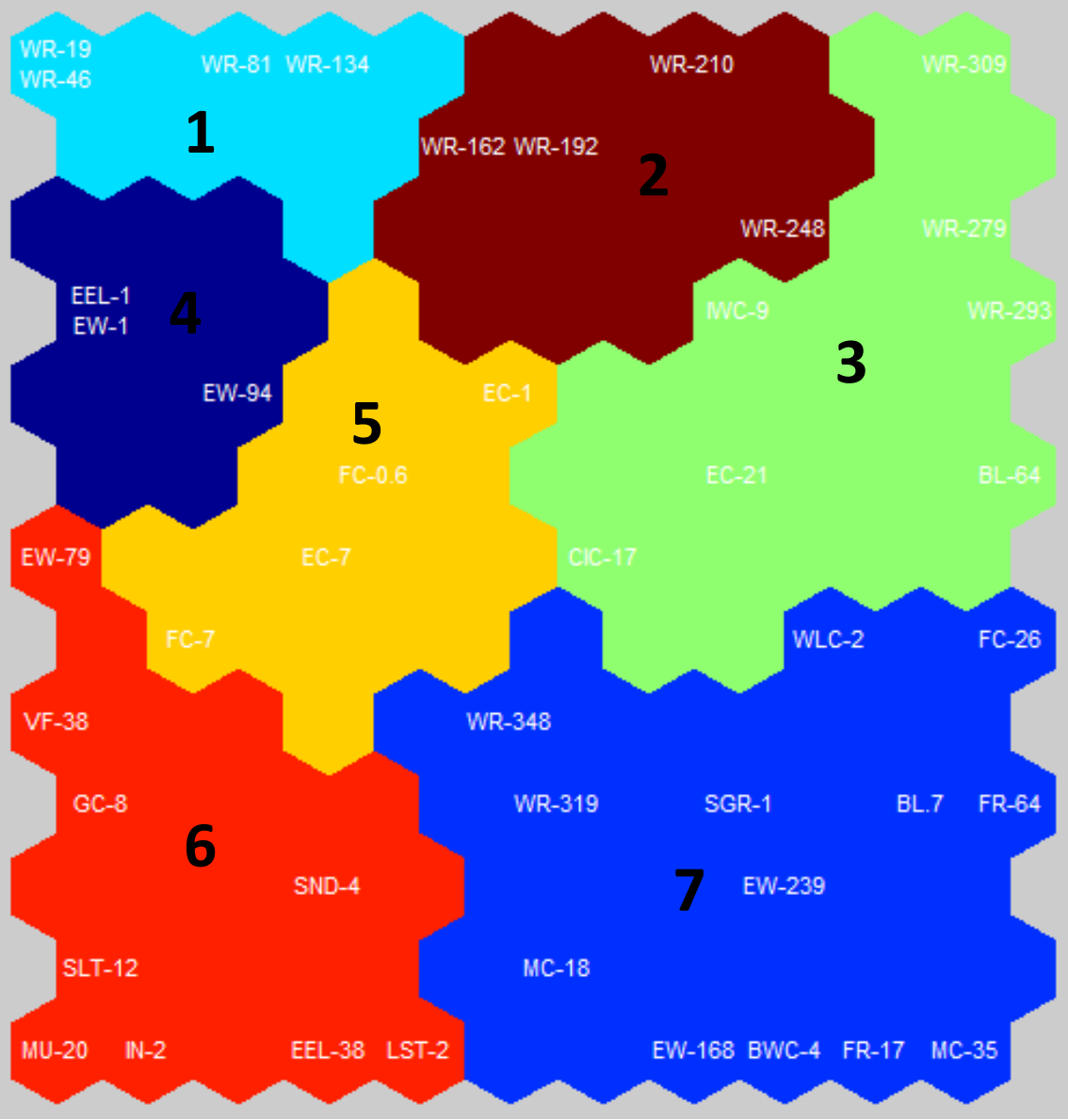
Supplementary Figure 4.13 – SOM cluster configuration for the quarter 3 mean dataset



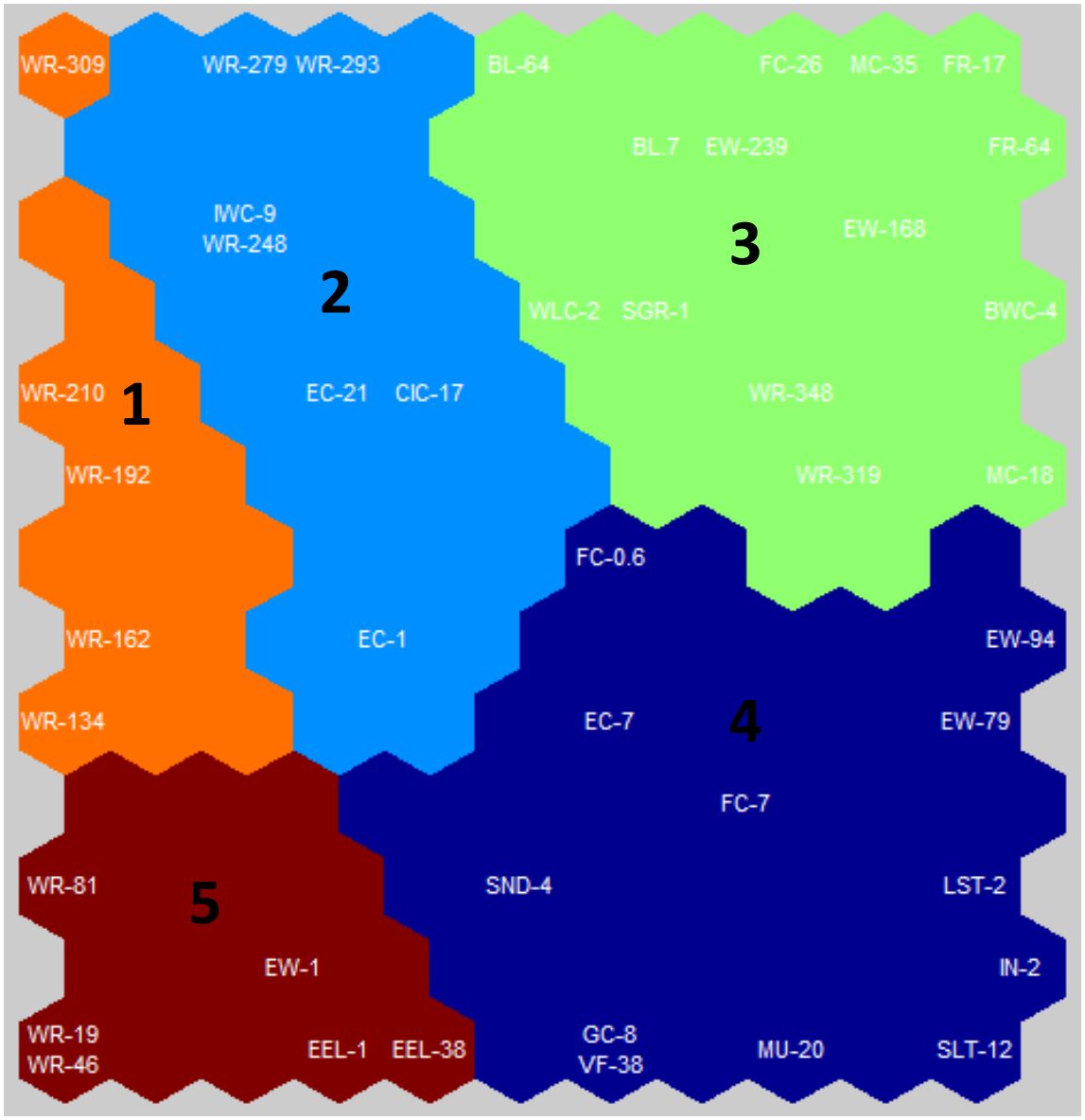
Supplementary Figure 4.14 – SOM cluster configuration for the quarter 3 median dataset



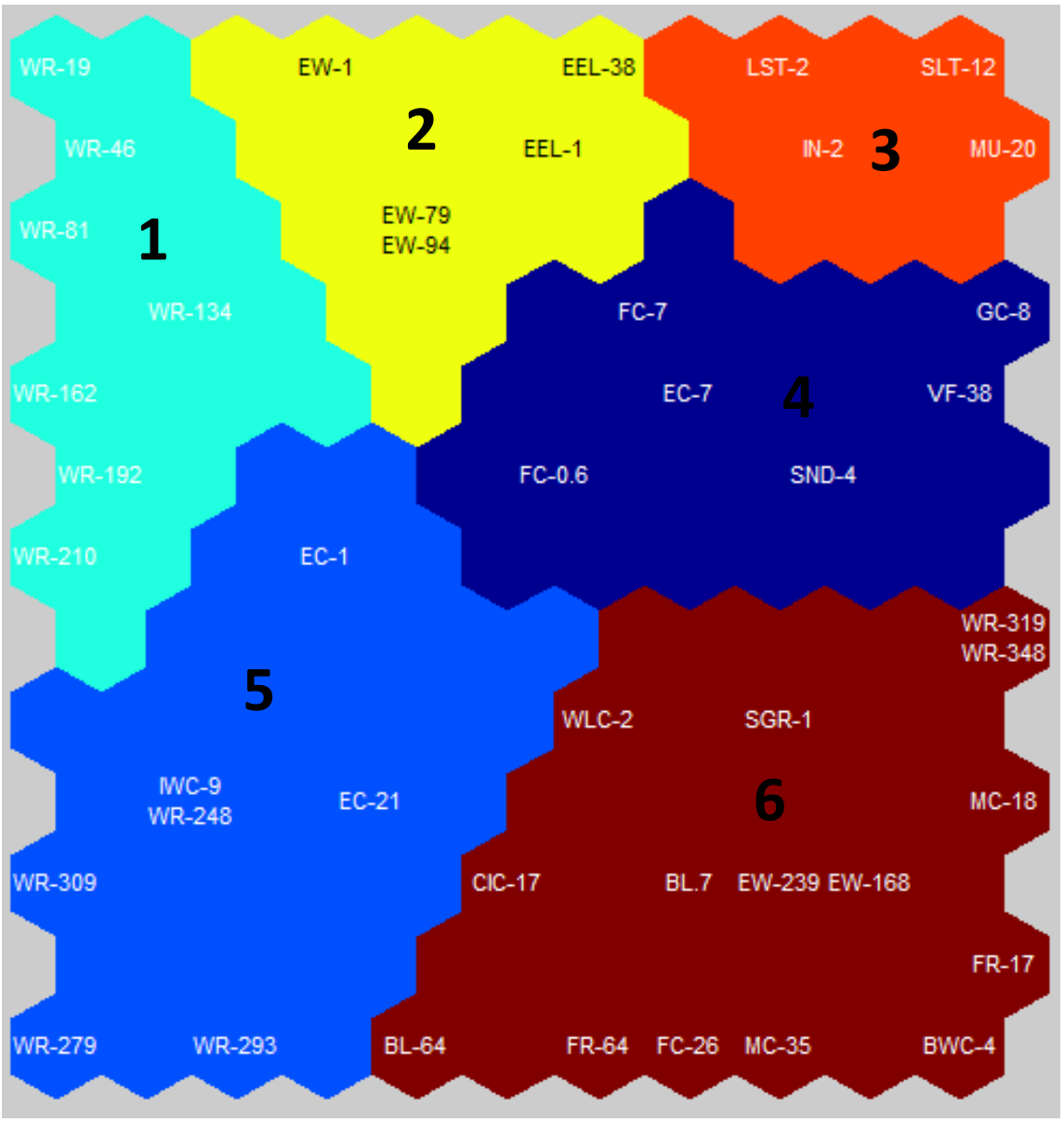
Supplementary Figure 4.15 – SOM cluster configuration for the quarter 3 trimmed mean dataset



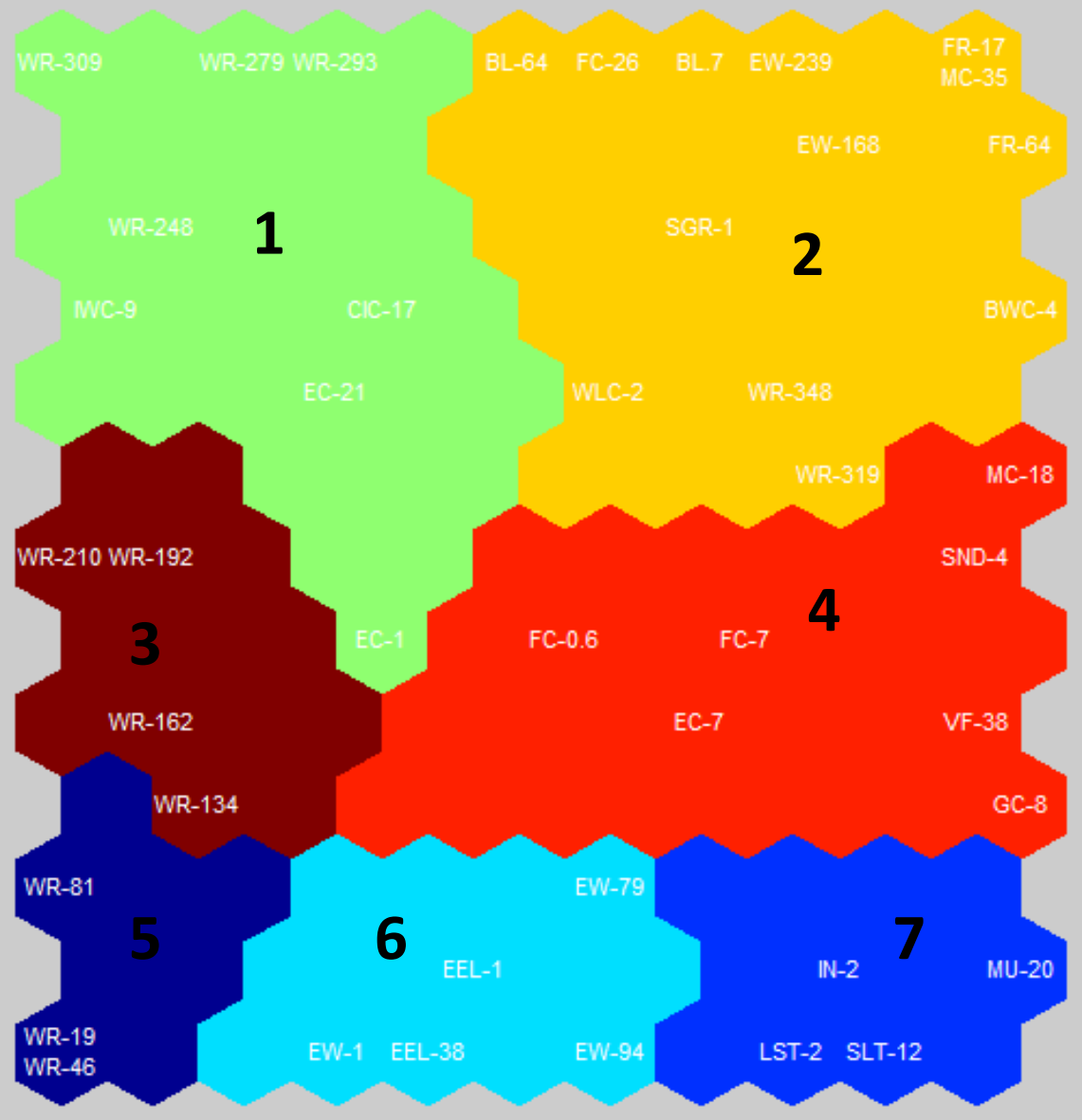
Supplementary Figure 4.16 – SOM cluster configuration for the quarter 3 geometric mean dataset



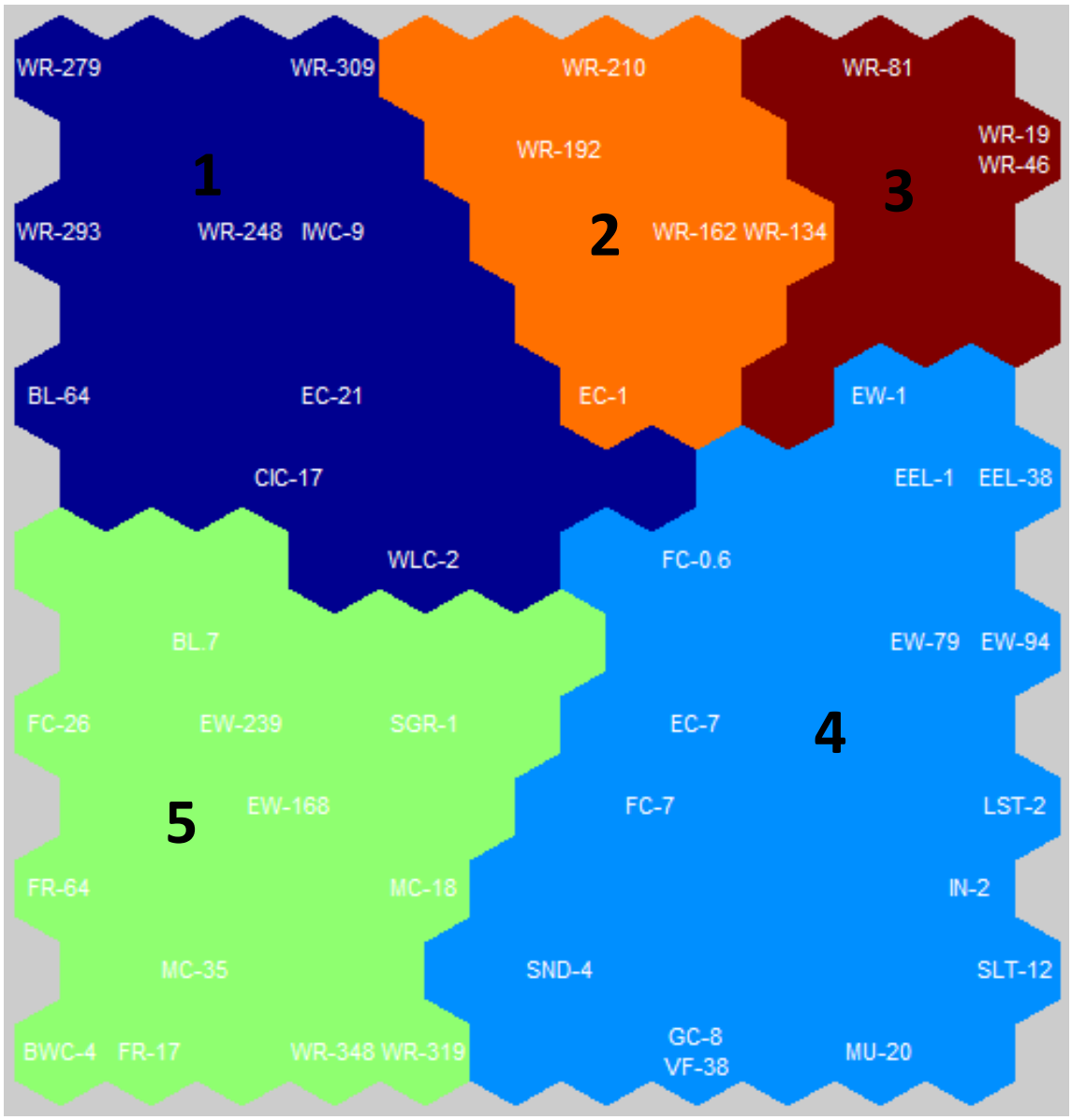
Supplementary Figure 4.17 – SOM cluster configuration for the quarter 4 mean dataset



Supplementary Figure 4.18 – SOM cluster configuration for the quarter 4 median dataset

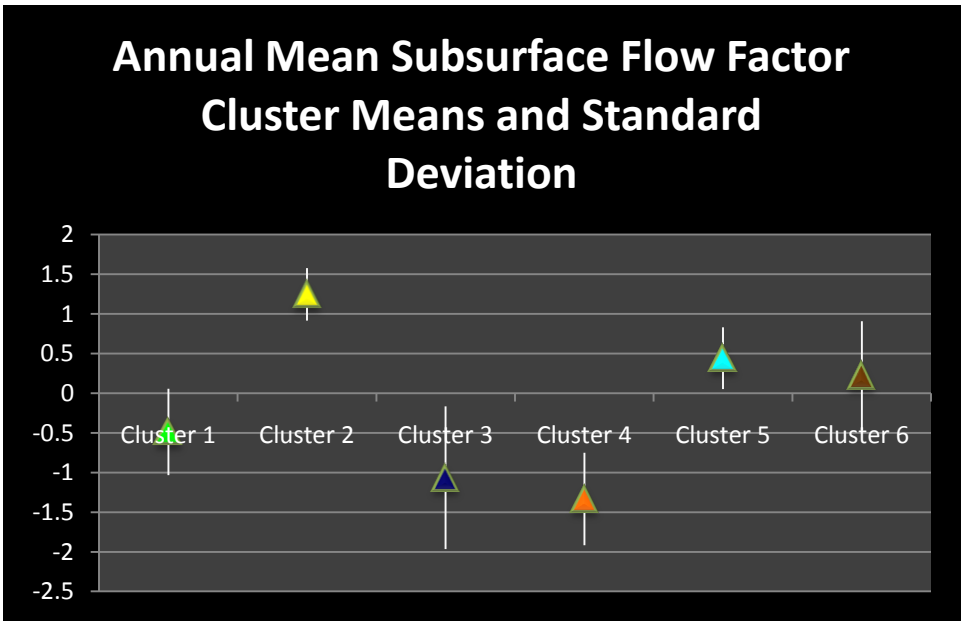


Supplementary Figure 4.19 – SOM cluster configuration for the quarter 4 trimmed mean dataset

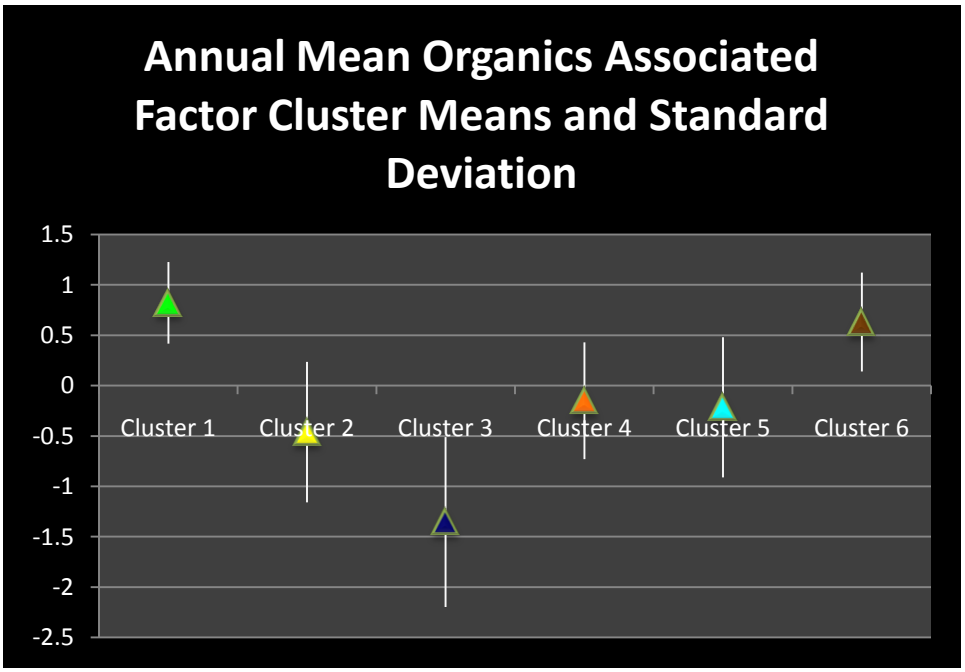


Supplementary Figure 4.20 – SOM cluster configuration for the quarter 4 geometric mean dataset

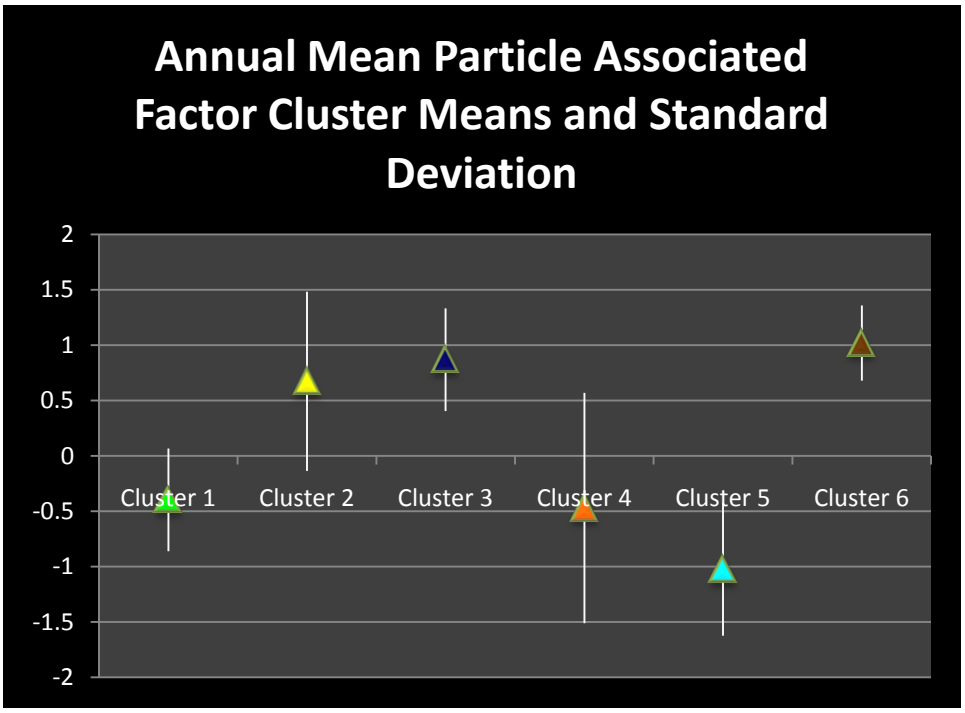
Mean and Standard Deviation Box-Plots of Factor Clusters



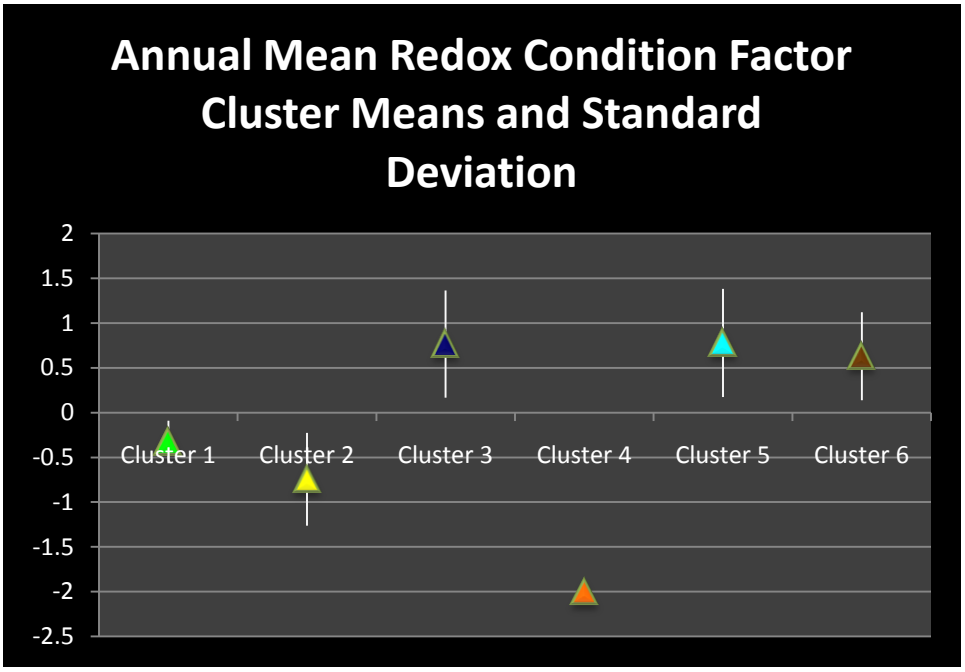
Supplementary Figure 5.1 – Cluster mean comparison for the subsurface flow associated factor for the annual mean dataset



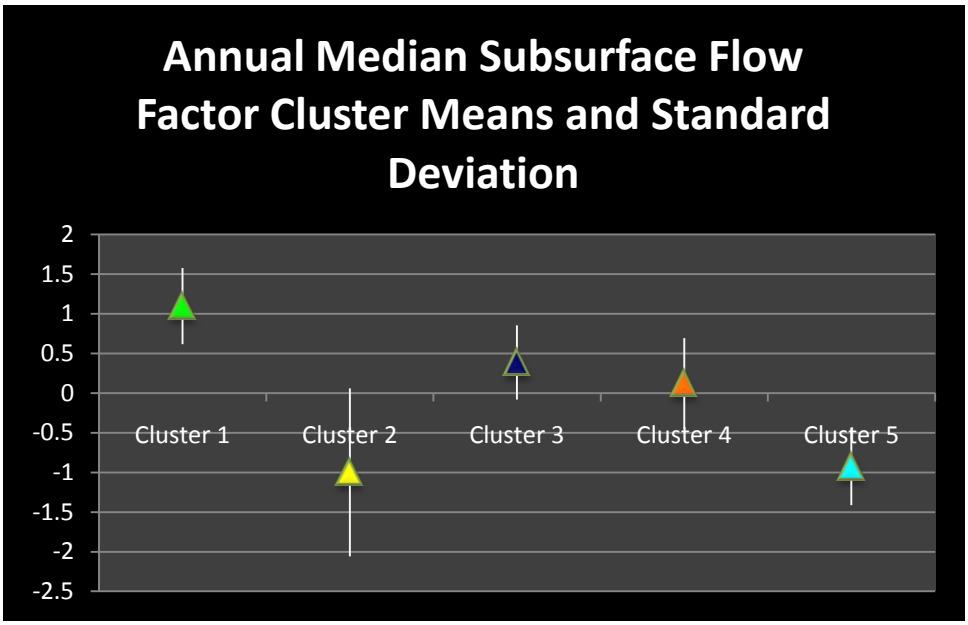
Supplementary Figure 5.2 – Cluster mean comparison for the organic associated factor for the annual mean dataset



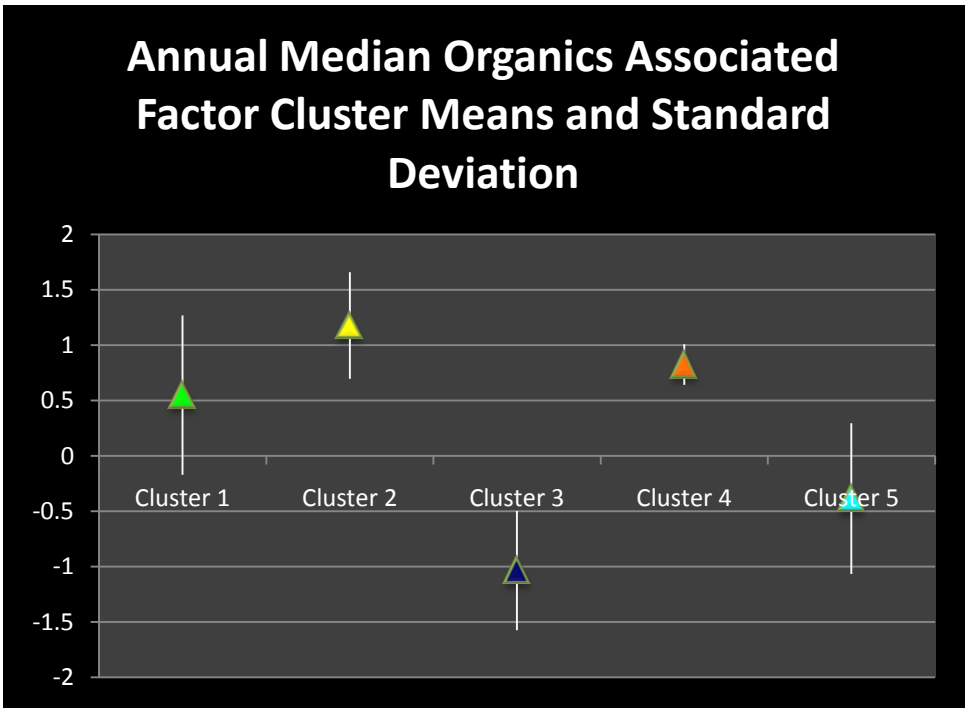
Supplementary Figure 5.3 – Cluster mean comparison for the particle associated factor for the annual mean dataset



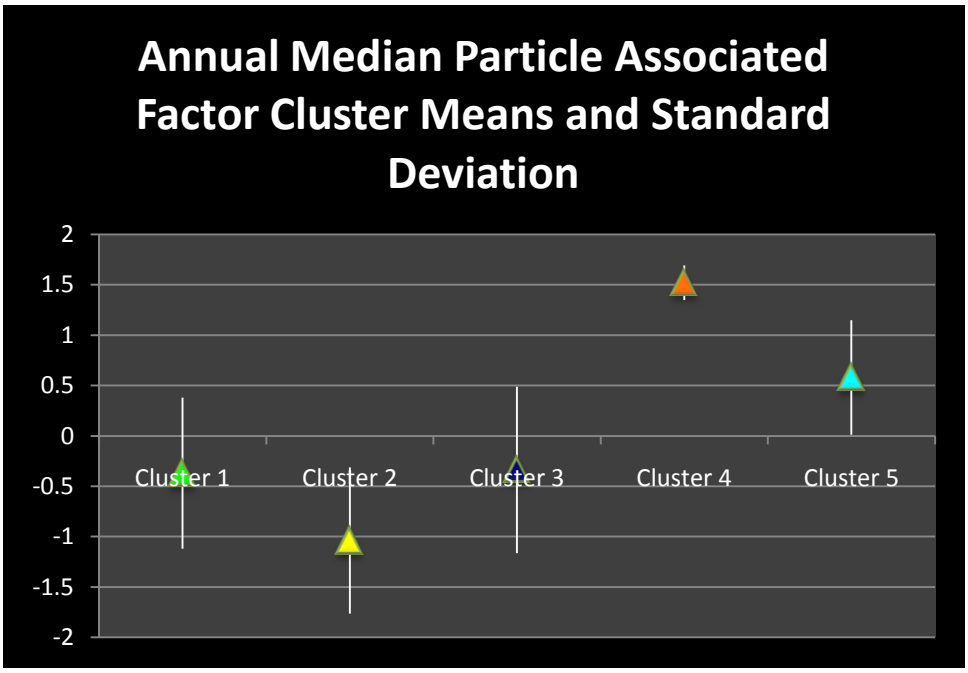
Supplementary Figure 5.4 – Cluster mean comparison for the redox conditions factor for the annual mean dataset



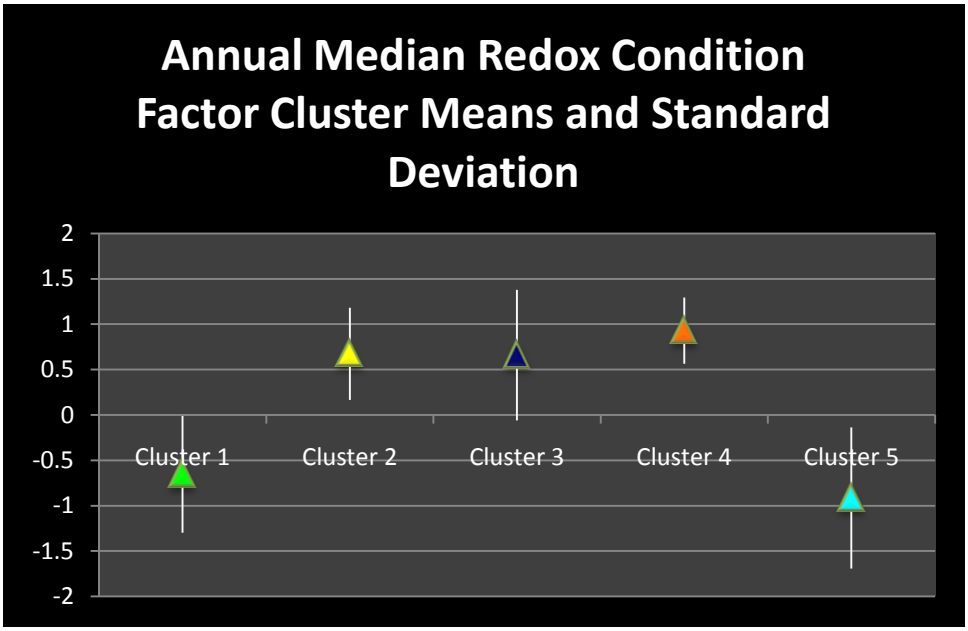
Supplementary Figure 5.5 – Cluster mean comparison for the subsurface flow associated factor for the annual median dataset



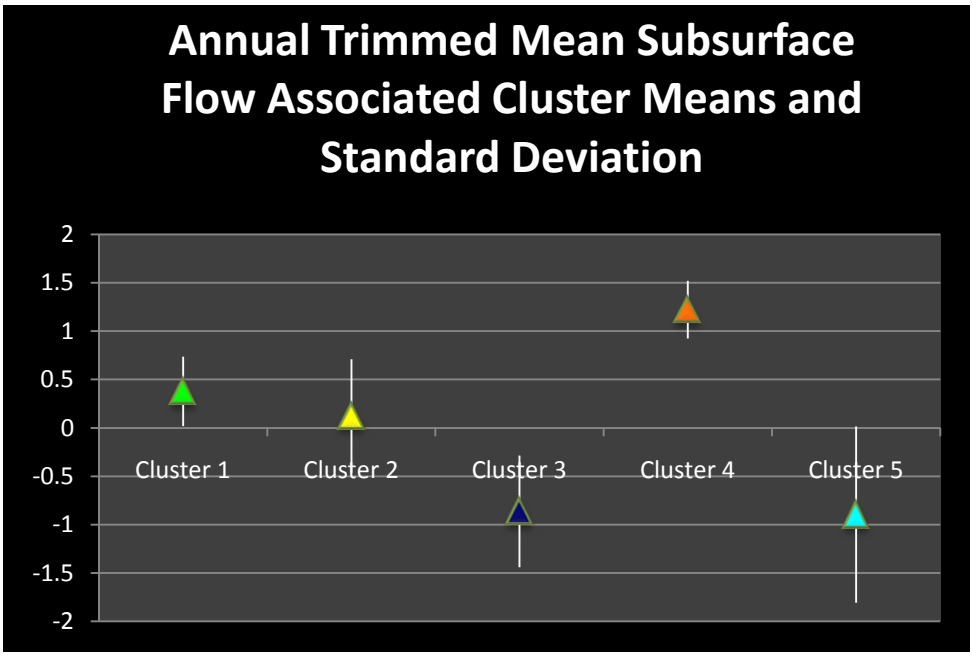
Supplementary Figure 5.6 – Cluster mean comparison for the organics associated factor for the annual median dataset



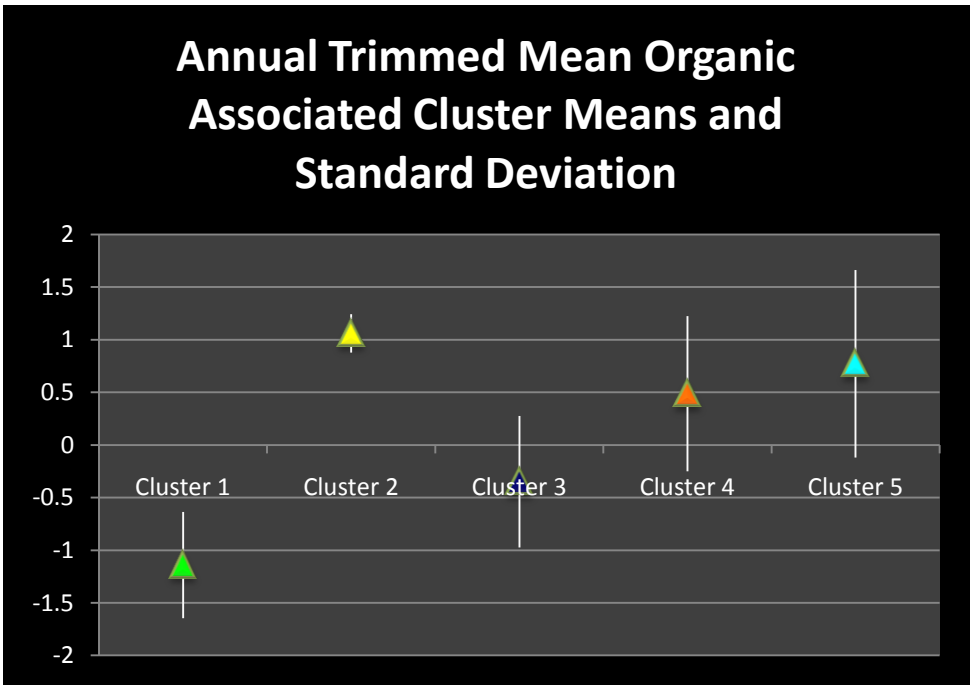
Supplementary Figure 5.7 – Cluster mean comparison for the particle associated factor for the annual median dataset



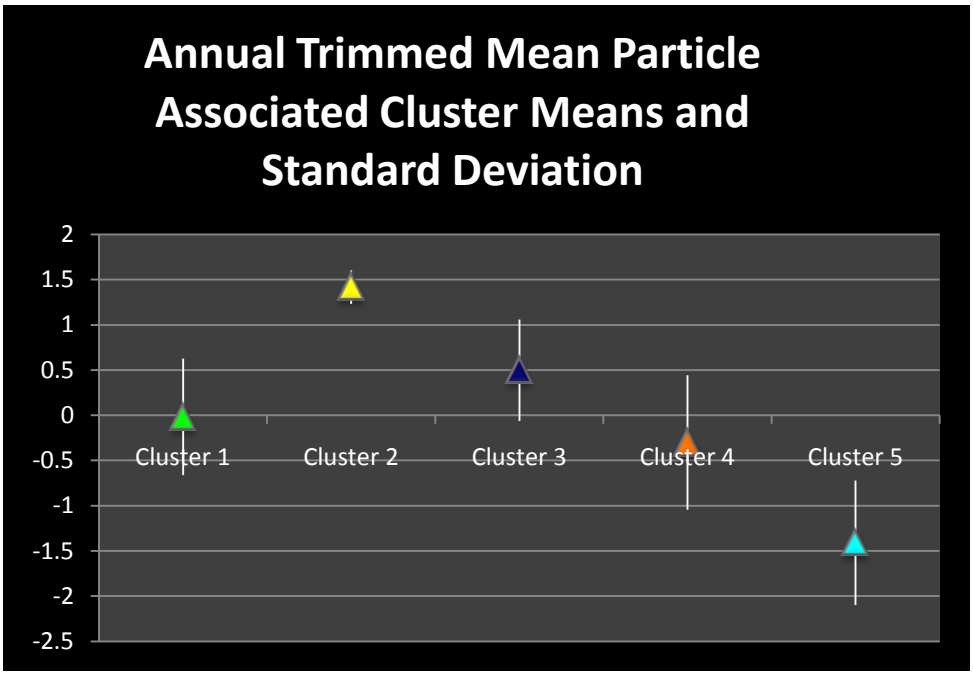
Supplementary Figure 5.8 – Cluster mean comparison for the redox conditions factor for the annual median dataset



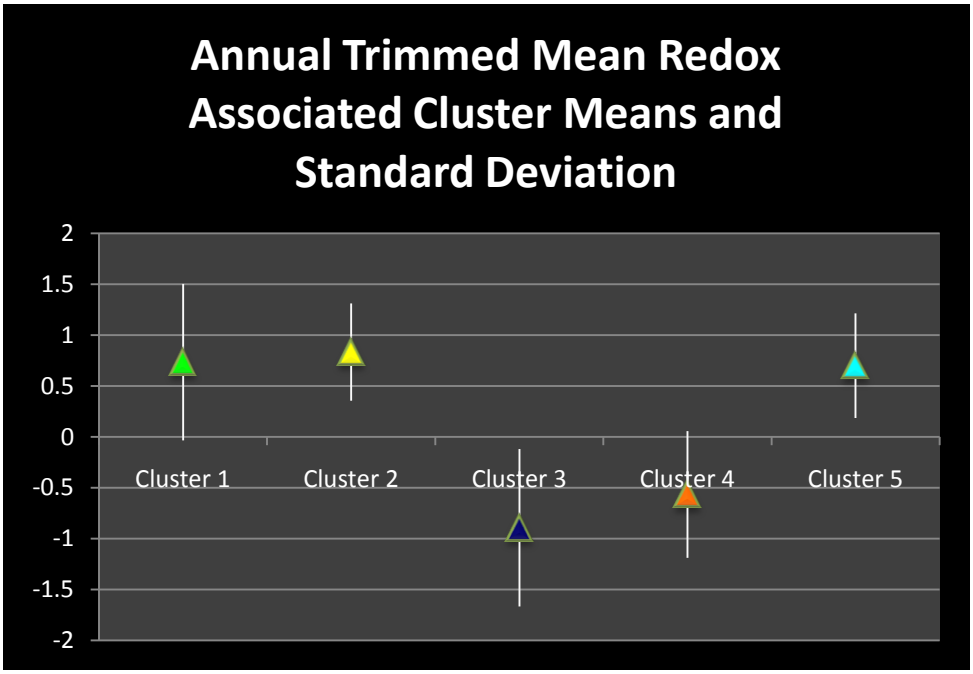
Supplementary Figure 5.9 – Cluster mean comparison for the subsurface flow associated factor for the annual trimmed mean dataset



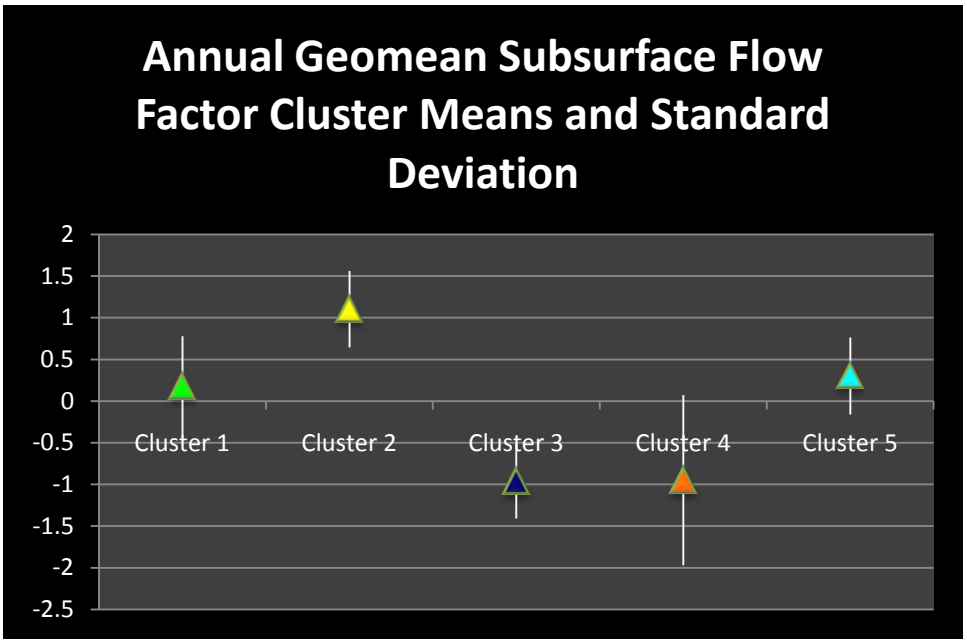
Supplementary Figure 5.10 – Cluster mean comparison for the organics associated factor for the annual trimmed mean dataset



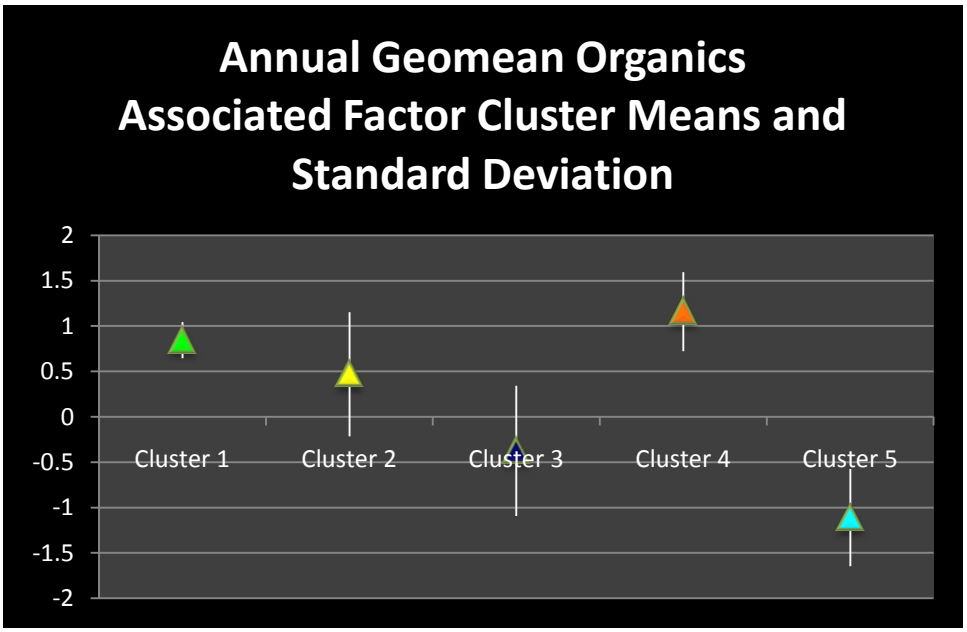
Supplementary Figure 5.11 – Cluster mean comparison for the particle associated factor for the annual trimmed mean dataset



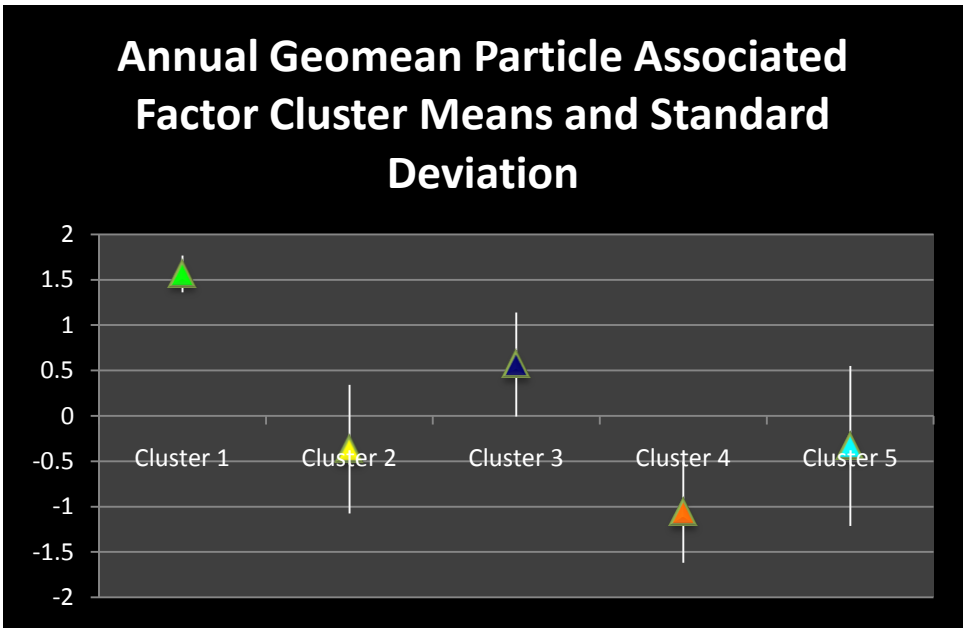
Supplementary Figure 5.12 – Cluster mean comparison for the redox conditions factor for the annual trimmed mean dataset



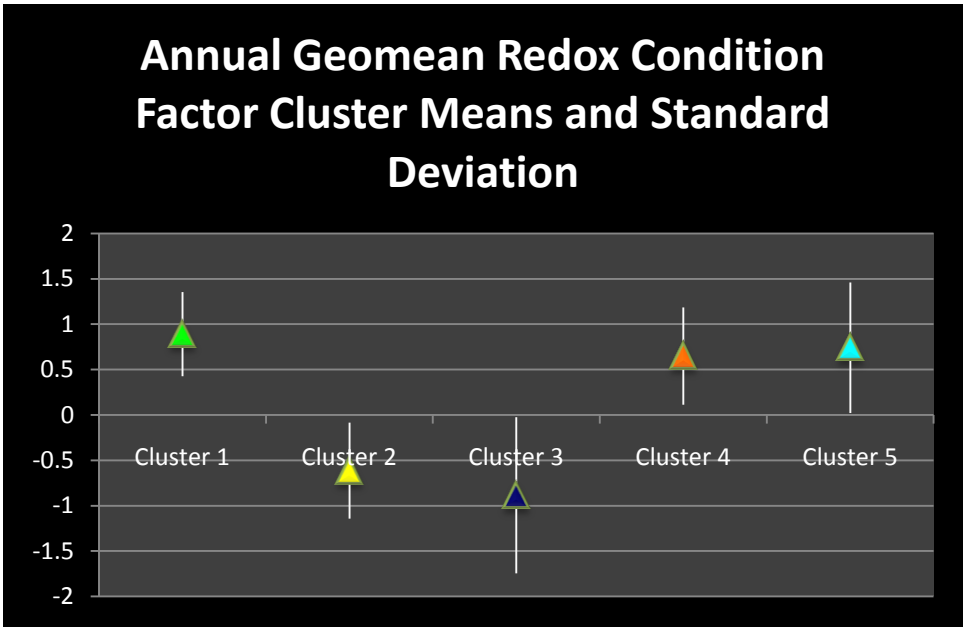
Supplementary Figure 5.13 – Cluster mean comparison for the subsurface flow associated factor for the annual geometric mean dataset



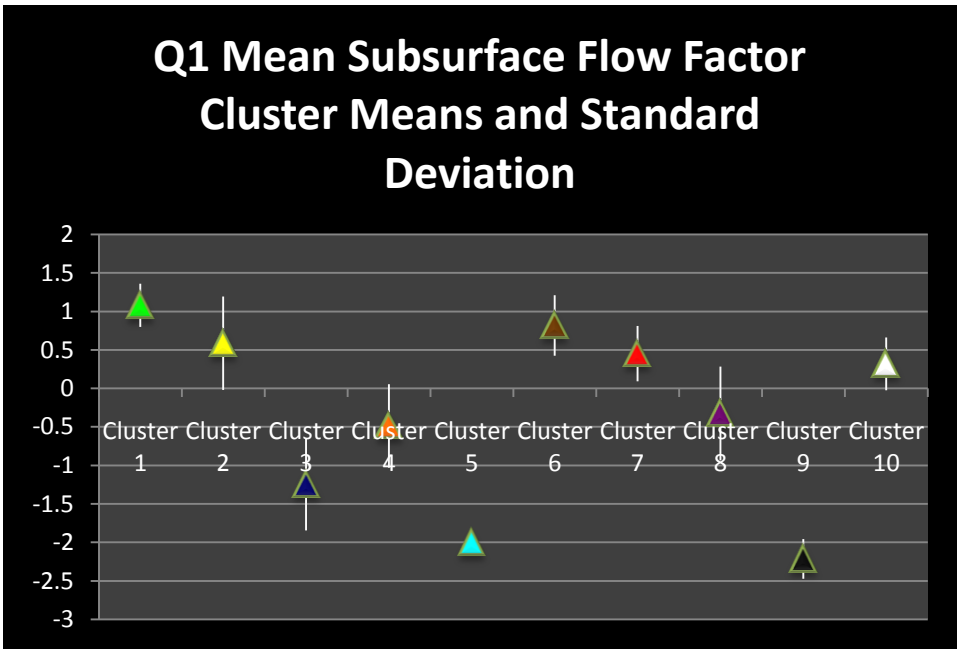
Supplementary Figure 5.14 – Cluster mean comparison for the organics associated factor for the annual geometric mean dataset



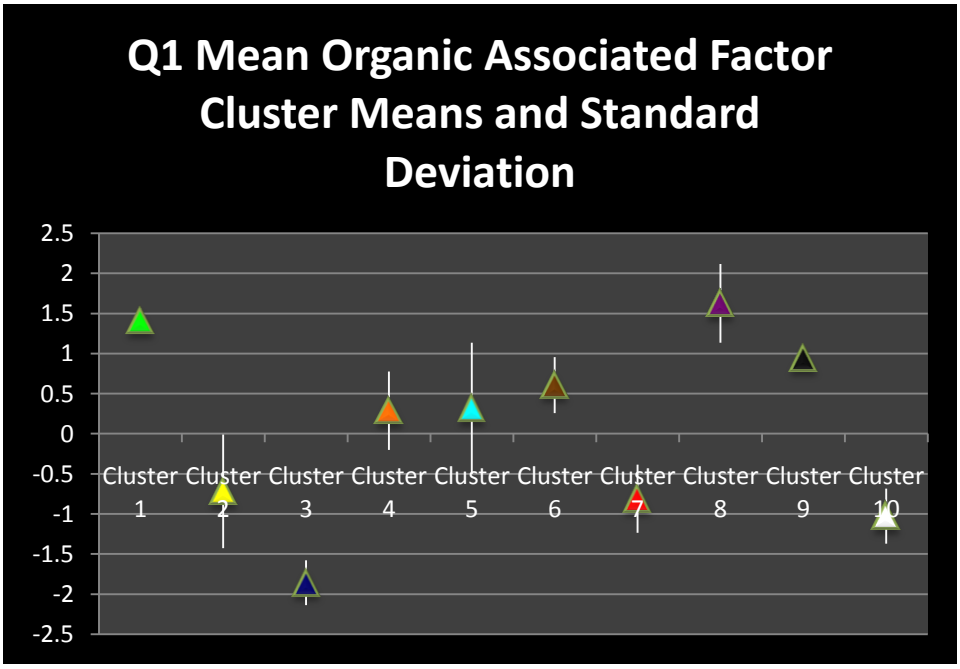
Supplementary Figure 5.15 – Cluster mean comparison for the particle associated factor for the annual geometric mean dataset



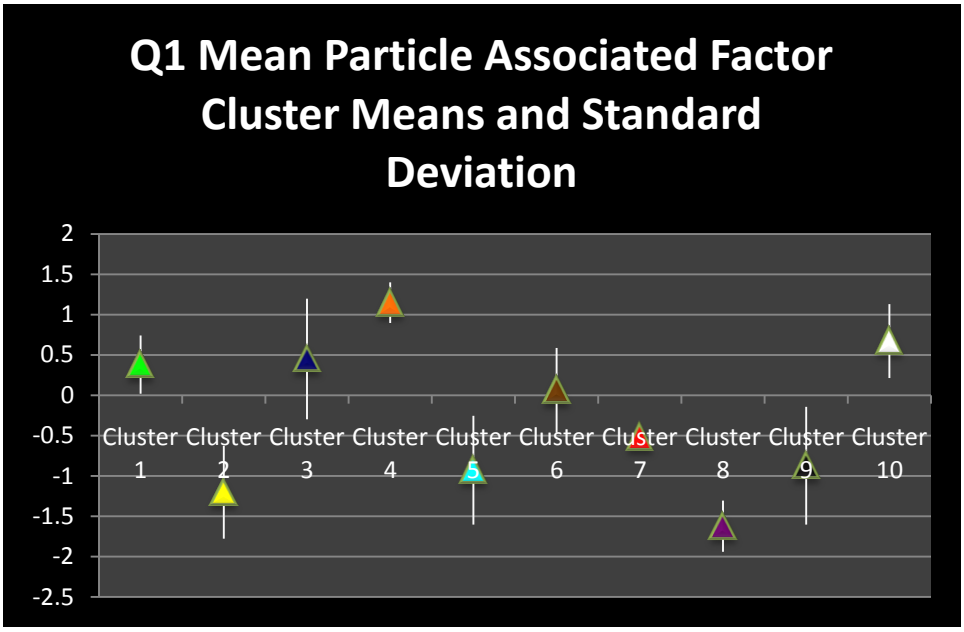
Supplementary Figure 5.16 – Cluster mean comparison for the redox conditions factor for the annual geometric mean dataset



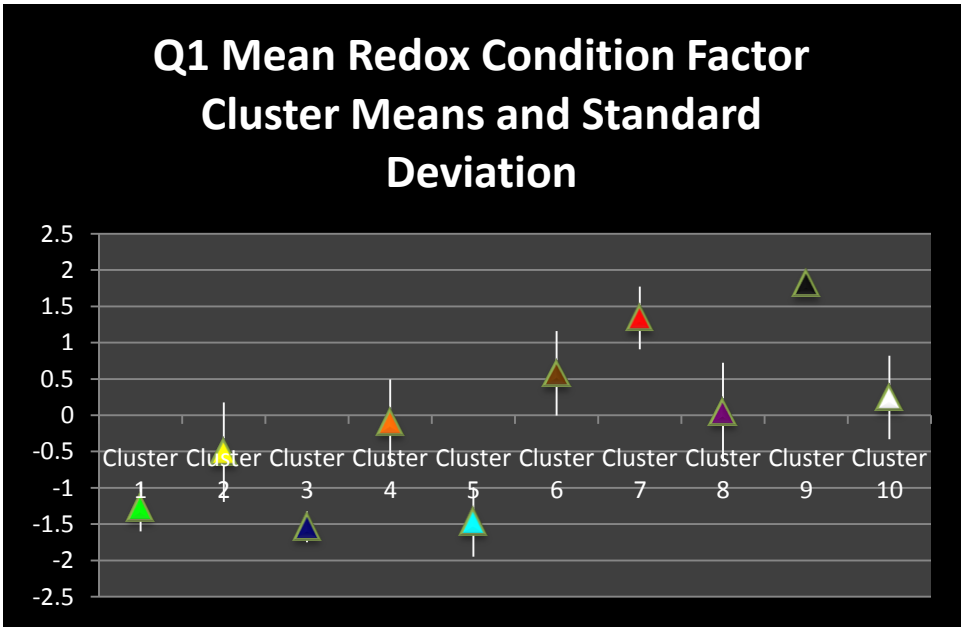
Supplementary Figure 5.17 – Cluster mean comparison for the subsurface flow associated factor for the quarter 1 mean dataset



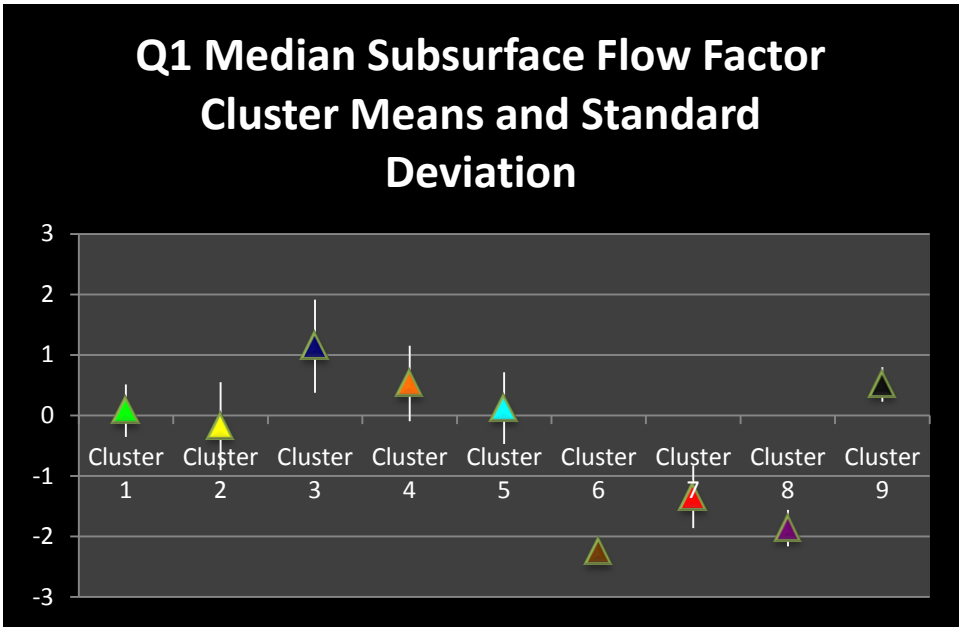
Supplementary Figure 5.18 – Cluster mean comparison for the organics associated factor for the quarter 1 mean dataset



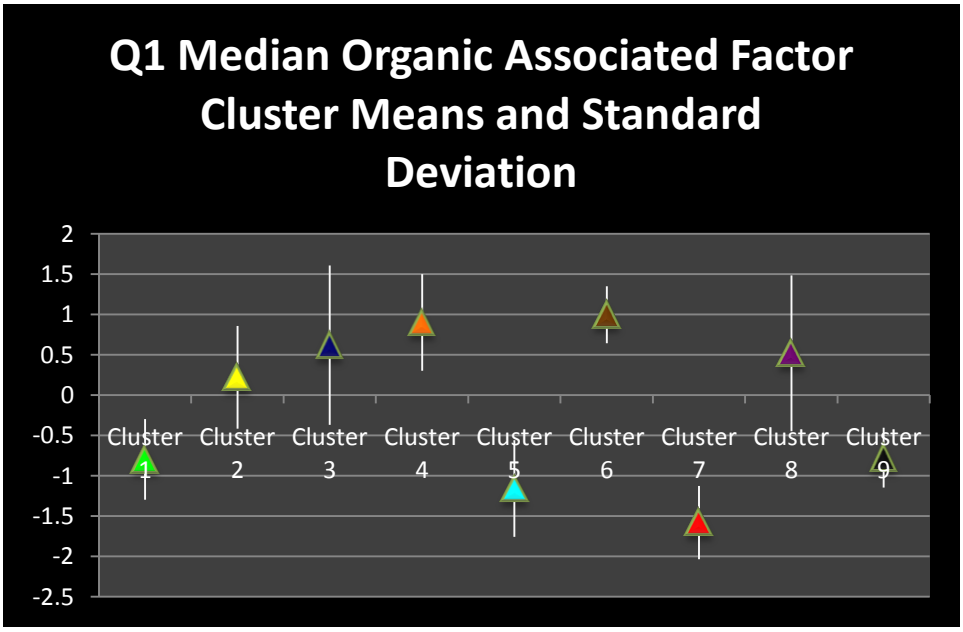
Supplementary Figure 5.19 – Cluster mean comparison for the particle associated factor for the quarter 1 mean dataset



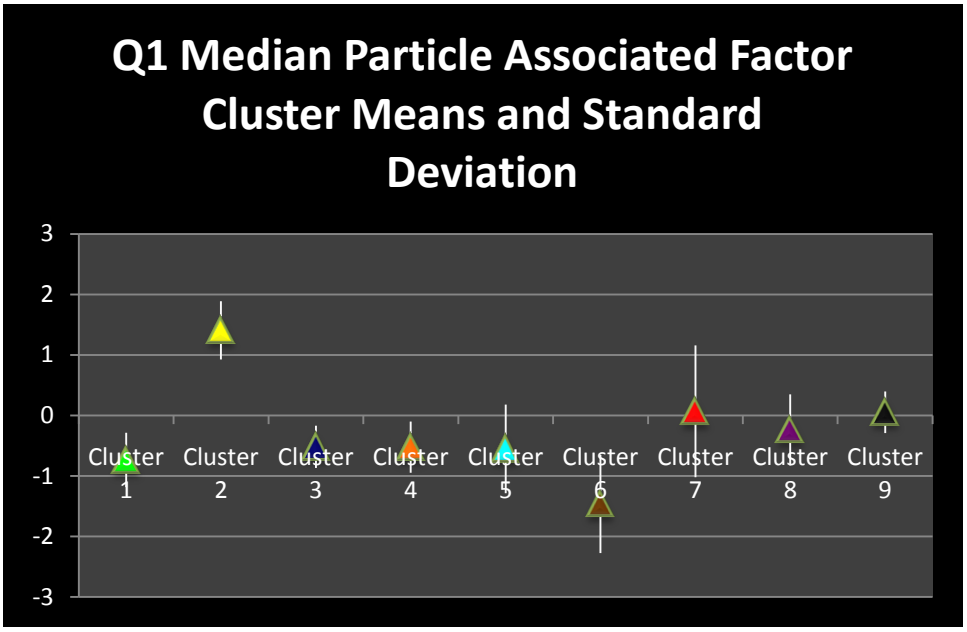
Supplementary Figure 5.20 – Cluster mean comparison for the redox conditions associated factor for the quarter 1 mean dataset



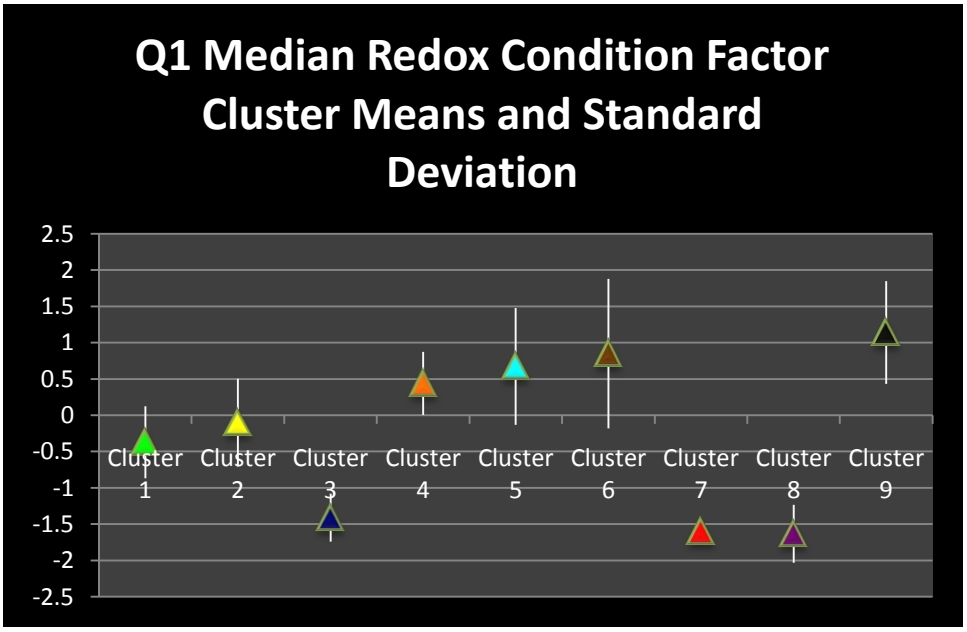
Supplementary Figure 5.21 – Cluster mean comparison for the subsurface flow associated factor for the quarter 1 median dataset



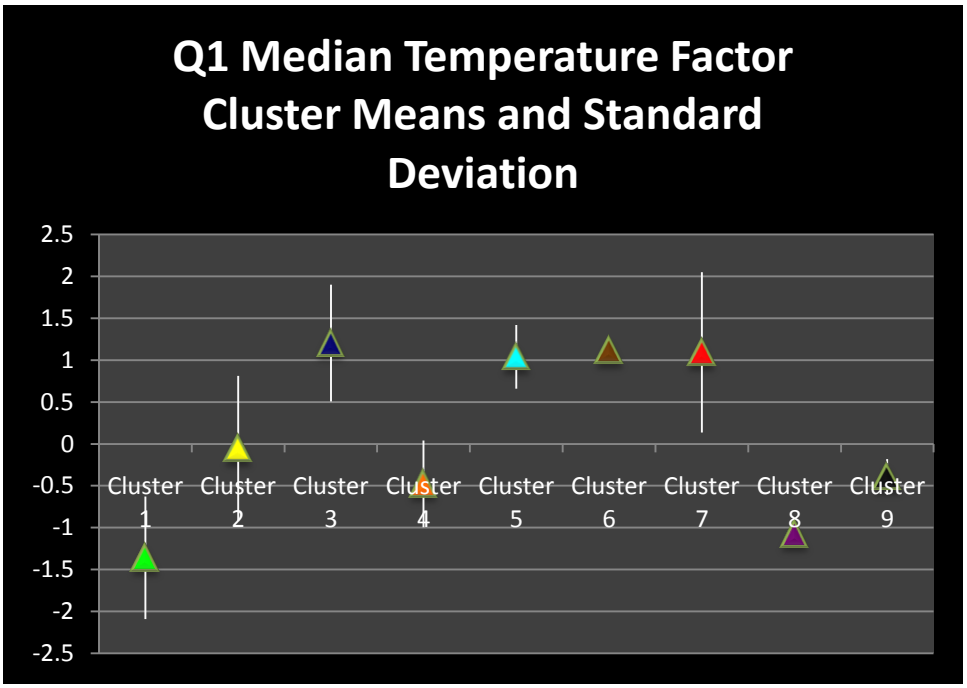
Supplementary Figure 5.22 – Cluster mean comparison for the organics associated factor for the quarter 1 median dataset



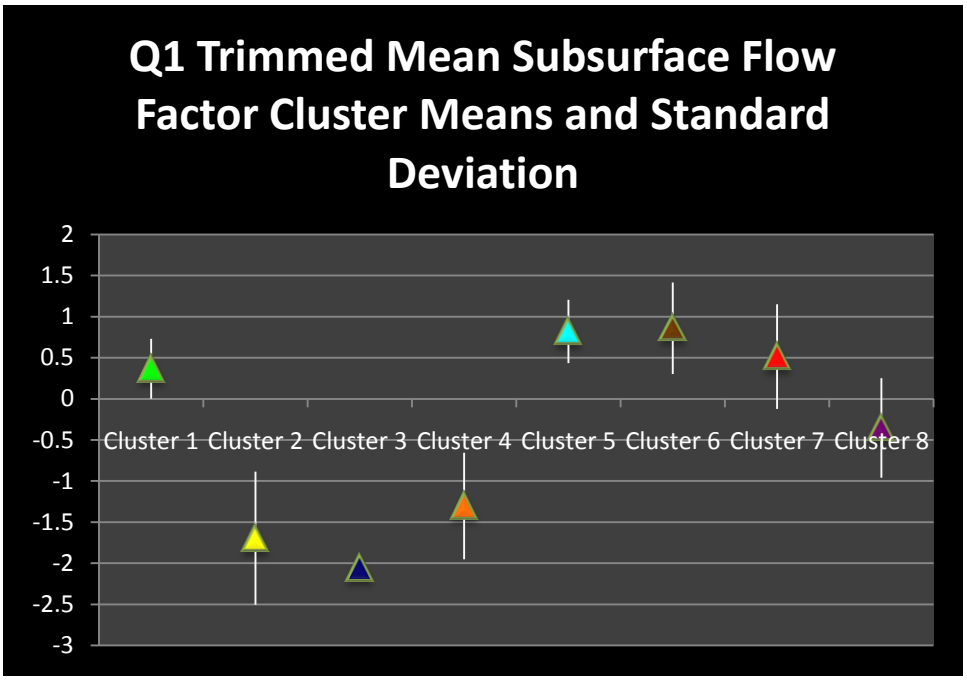
Supplementary Figure 5.23 – Cluster mean comparison for the particle associated factor for the quarter 1 median dataset



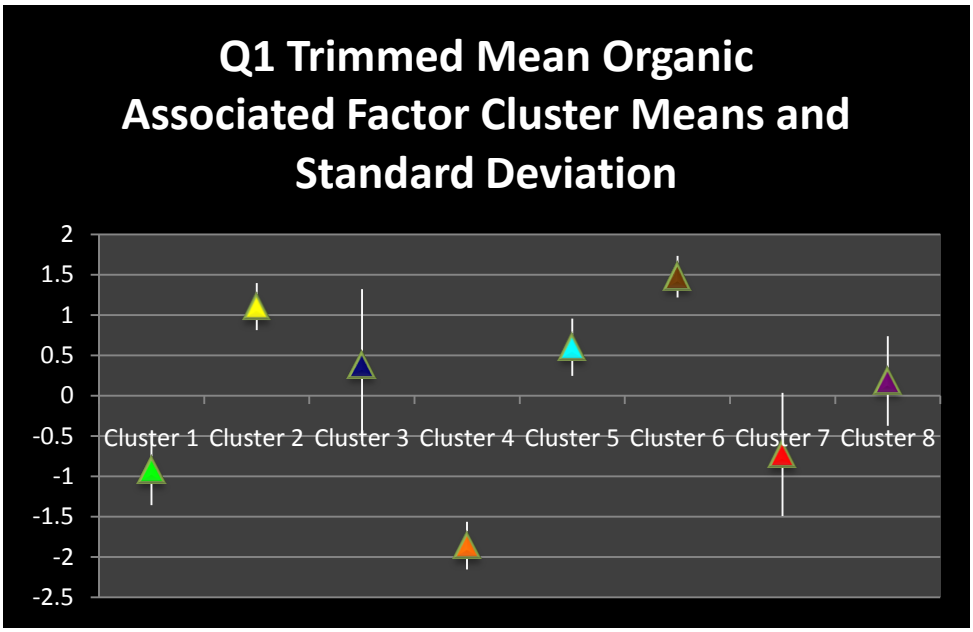
Supplementary Figure 5.24 – Cluster mean comparison for the redox conditions factor for the quarter 1 median dataset



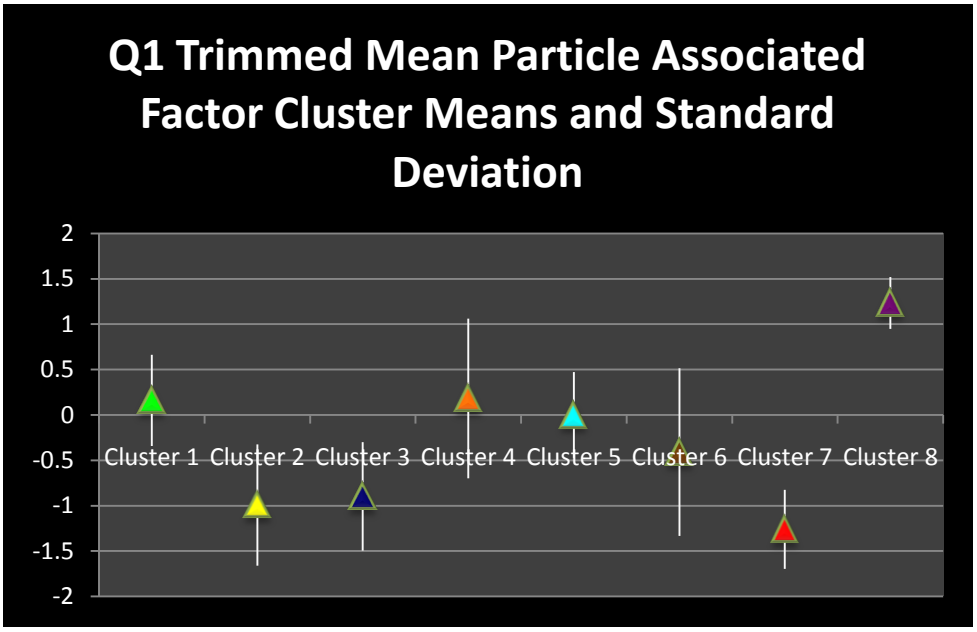
Supplementary Figure 5.25 – Cluster mean comparison for the temperature associated factor for the quarter 1 median dataset



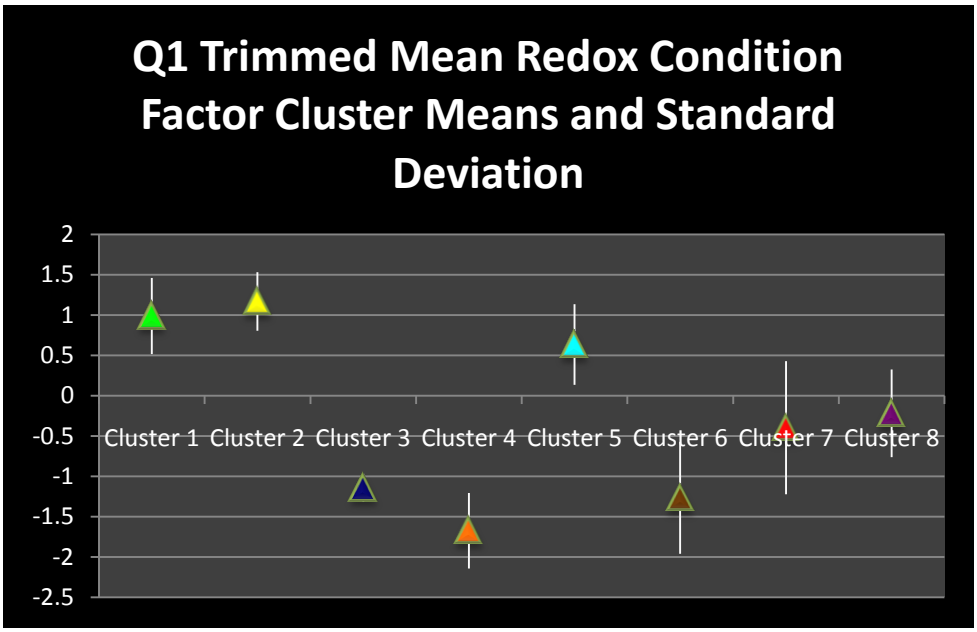
Supplementary Figure 5.26 – Cluster mean comparison for the subsurface flow associated factor for the quarter 1 trimmed mean dataset



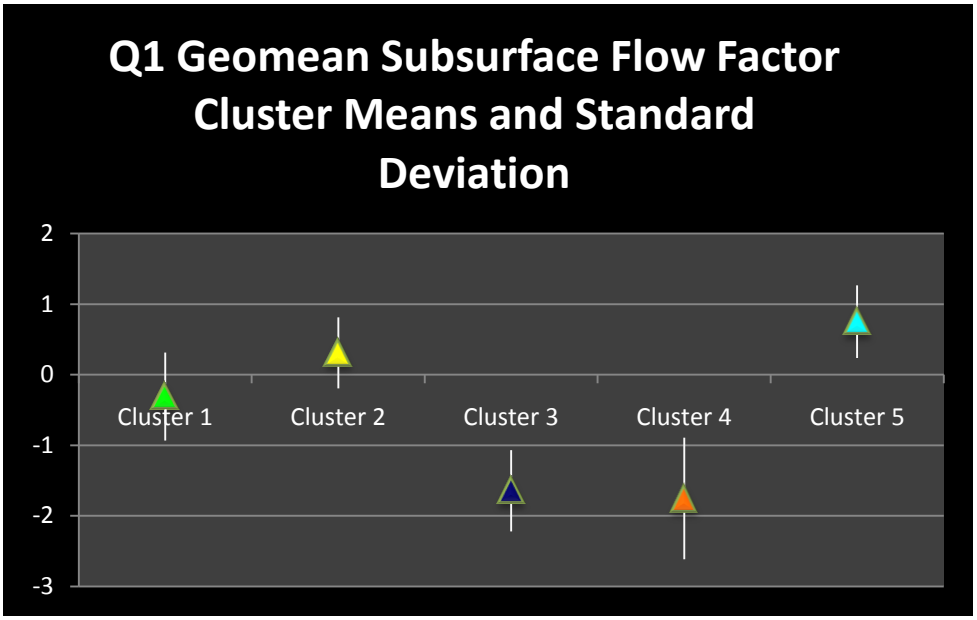
Supplementary Figure 5.27 – Cluster mean comparison for the organics associated factor for the quarter 1 trimmed mean dataset



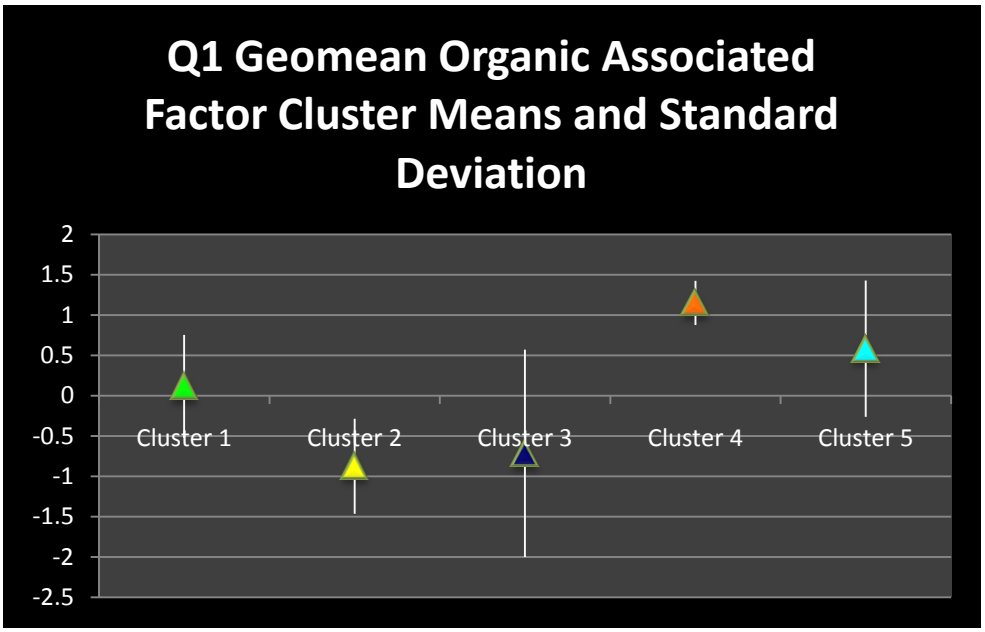
Supplementary Figure 5.28 – Cluster mean comparison for the particle associated factor for the quarter 1 trimmed mean dataset



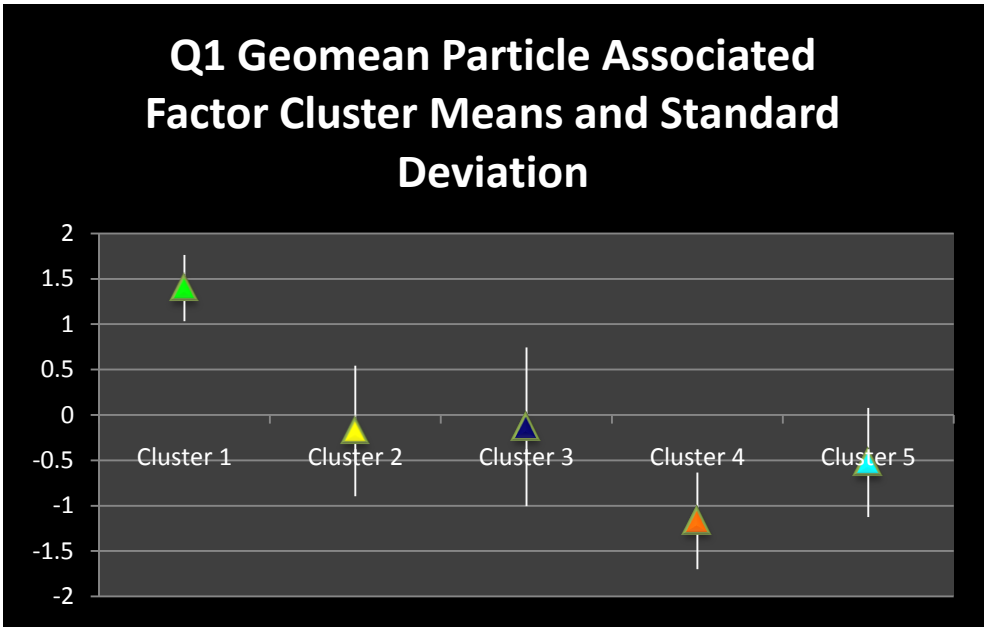
Supplementary Figure 5.29 – Cluster mean comparison for the redox conditions factor for the quarter 1 trimmed mean dataset



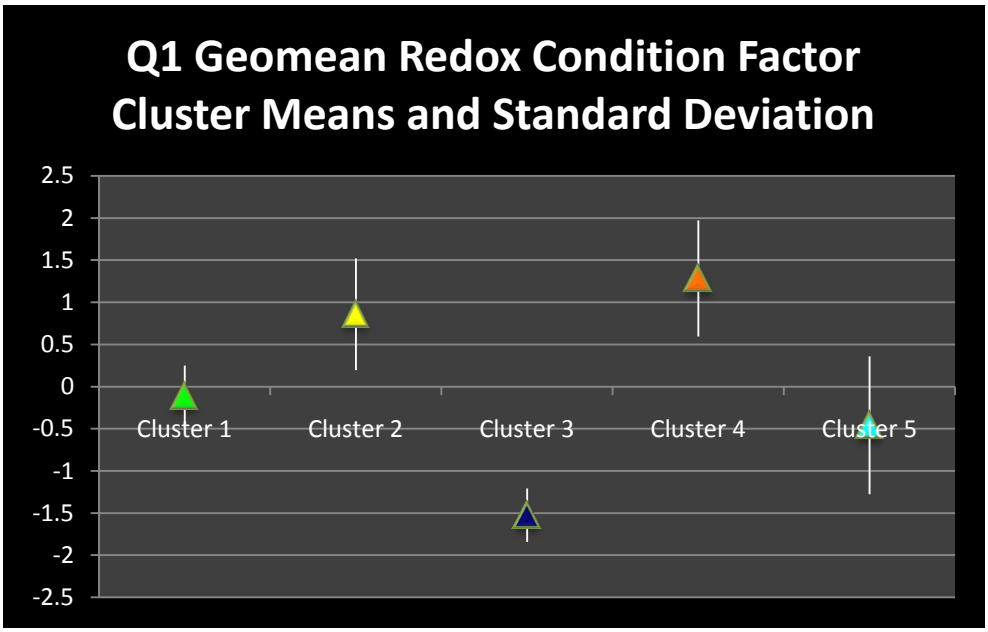
Supplementary Figure 5.30 – Cluster mean comparison for the subsurface flow associated factor for the quarter 1 geometric mean dataset



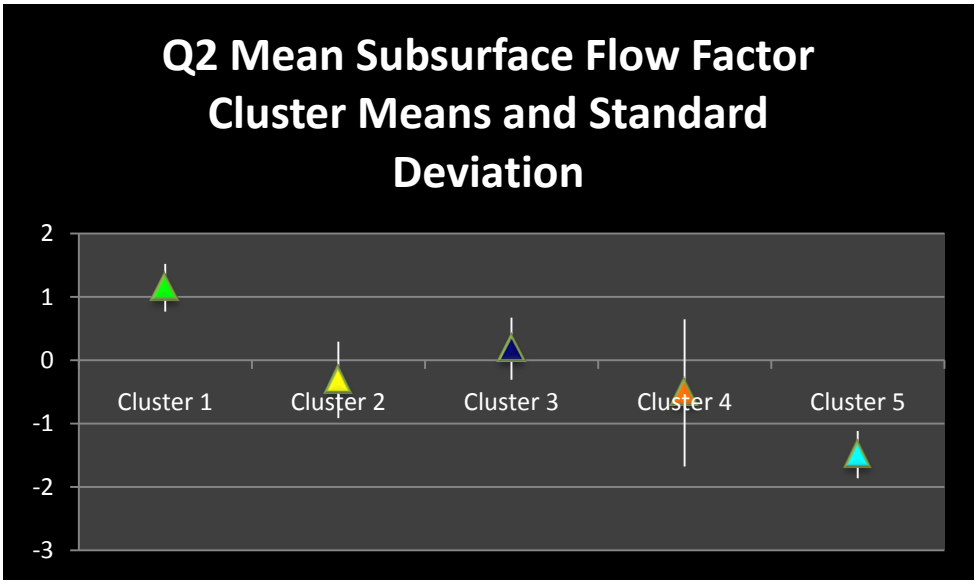
Supplementary Figure 5.31 – Cluster mean comparison for the organics associated factor for the quarter 1 geometric mean dataset



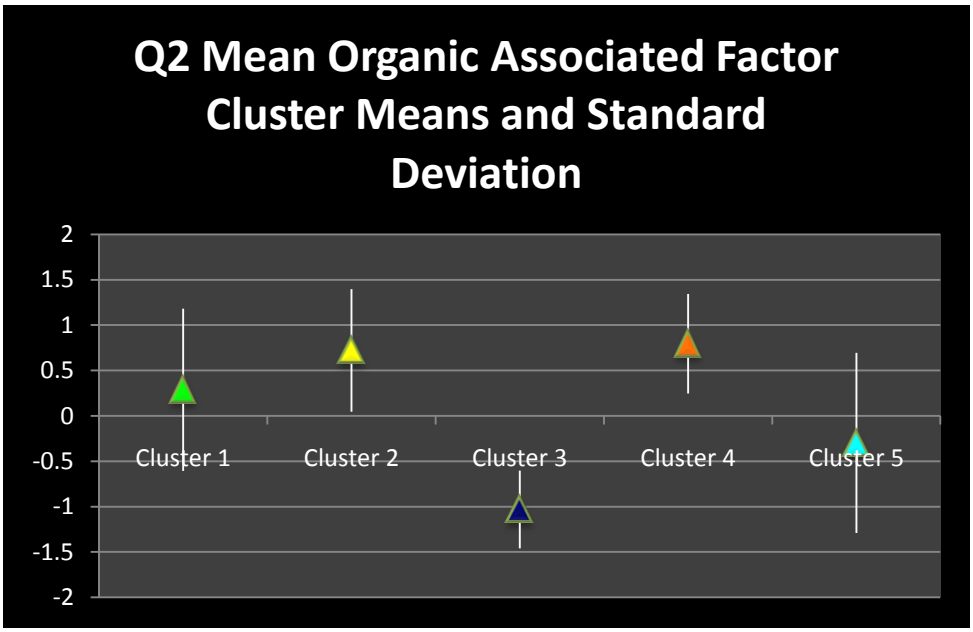
Supplementary Figure 5.32 – Cluster mean comparison for the particle associated factor for the quarter 1 geometric mean dataset



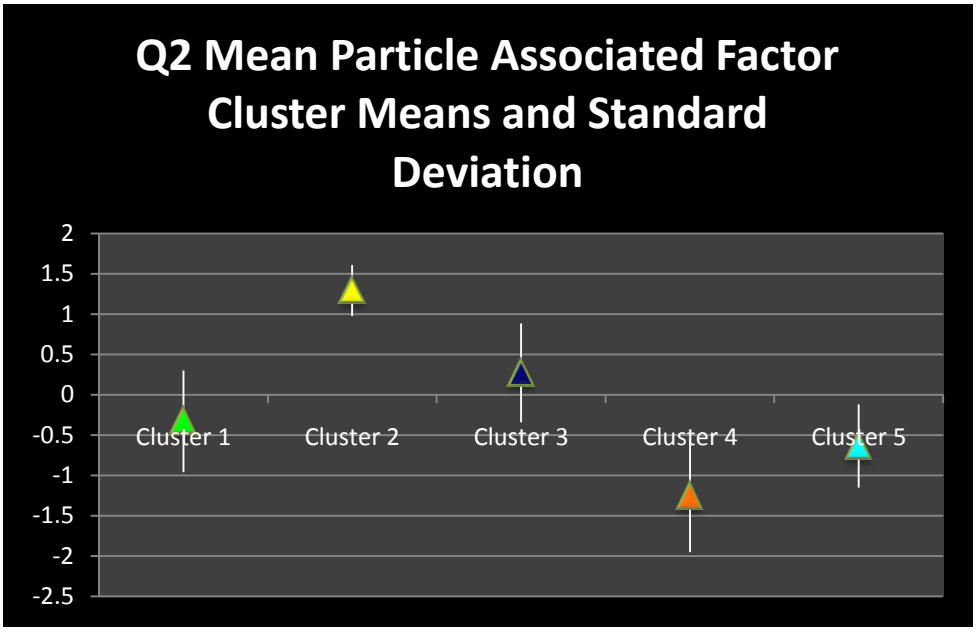
Supplementary Figure 5.33 – Cluster mean comparison for the redox conditions factor for the quarter 1 geometric mean dataset



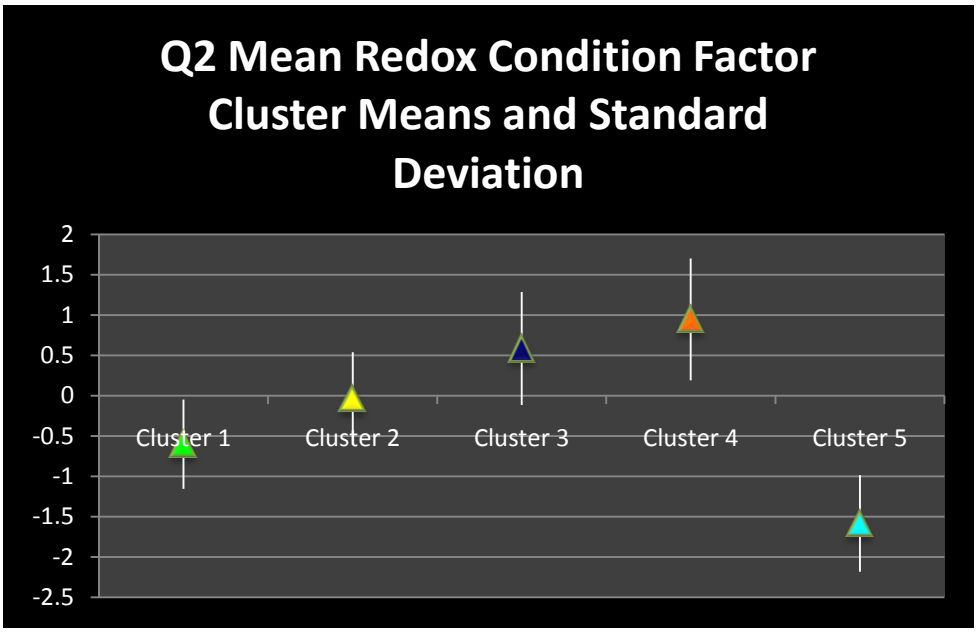
Supplementary Figure 5.34 – Cluster mean comparison for the subsurface flow associated factor for the quarter 2 mean dataset



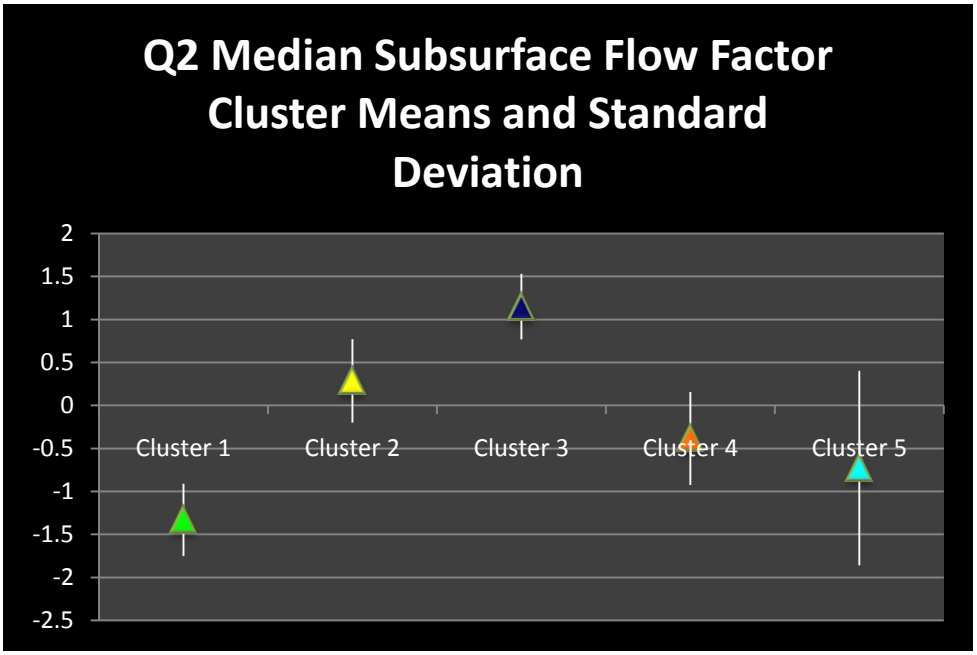
Supplementary Figure 5.35 – Cluster mean comparison for the organics associated factor for the quarter 2 mean dataset



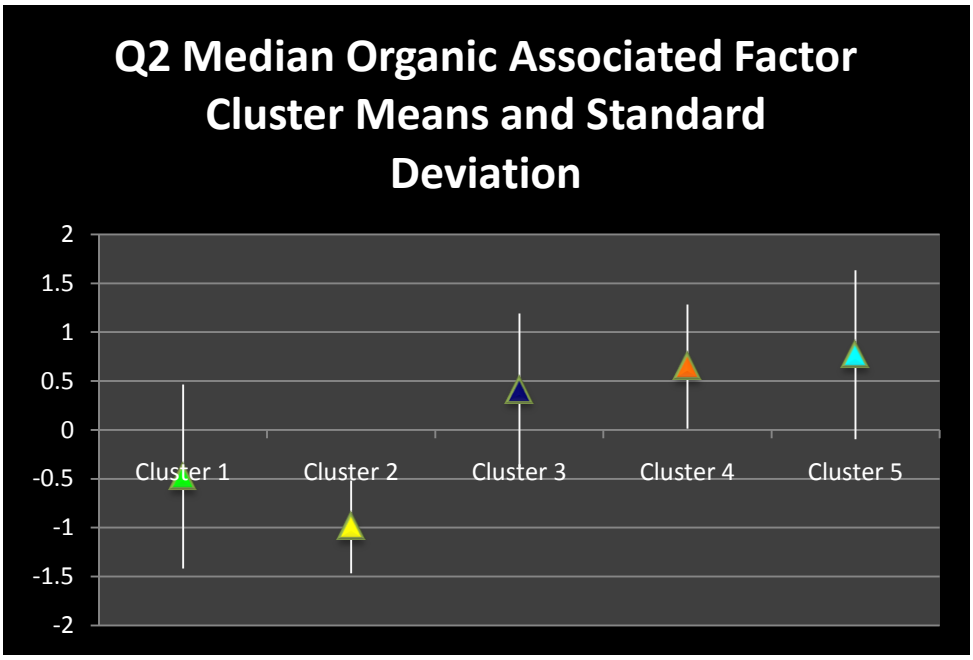
Supplementary Figure 5.36 – Cluster mean comparison for the particle associated factor for the quarter 2 mean dataset



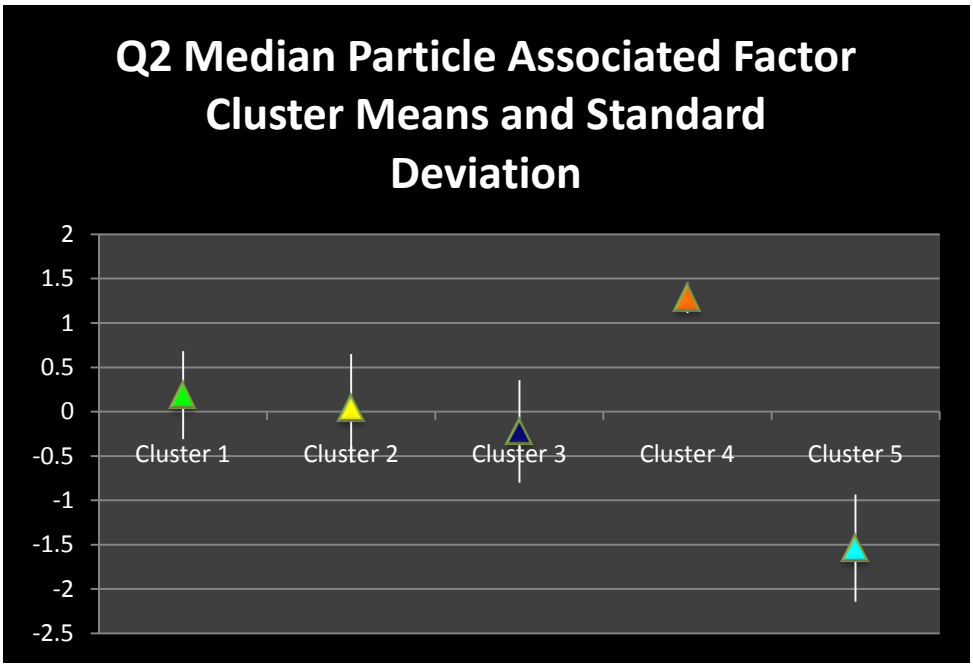
Supplementary Figure 5.37 – Cluster mean comparison for the redox conditions factor for the quarter 2 mean dataset



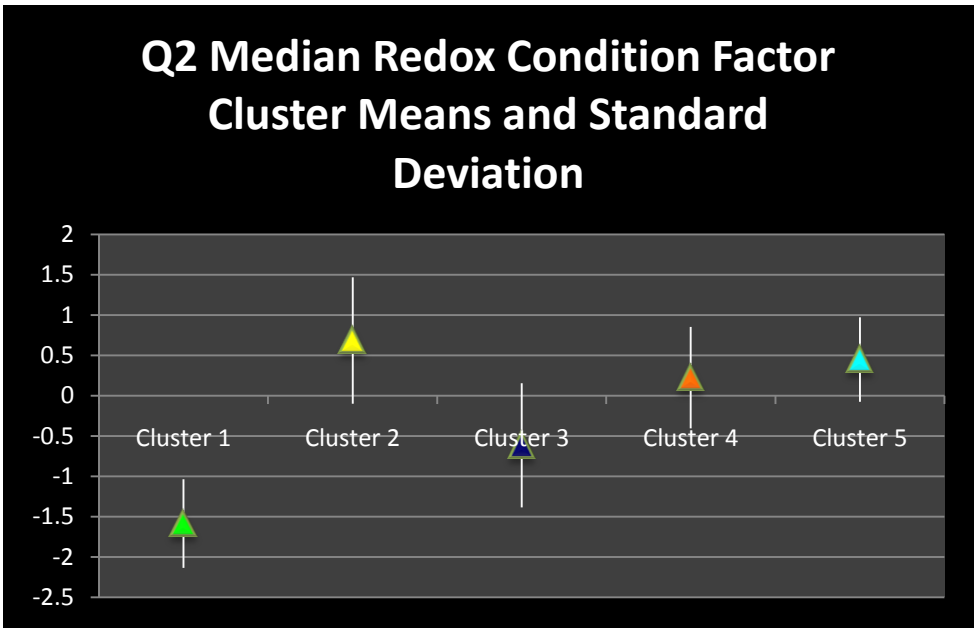
Supplementary Figure 5.38 – Cluster mean comparison for the subsurface flow associated factor for the quarter 2 median dataset



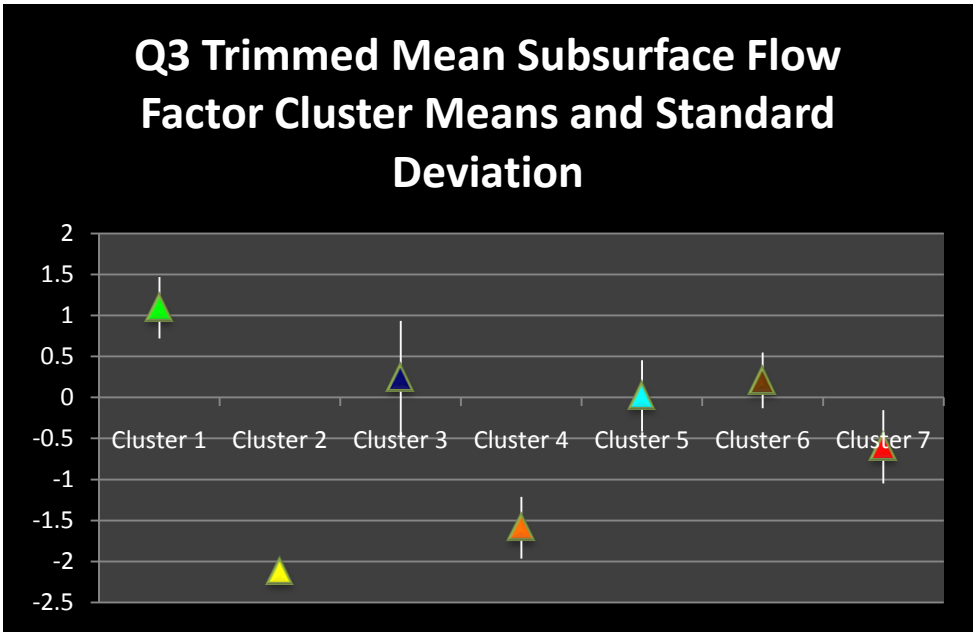
Supplementary Figure 5.39 – Cluster mean comparison for the organics associated factor for the quarter 2 median dataset



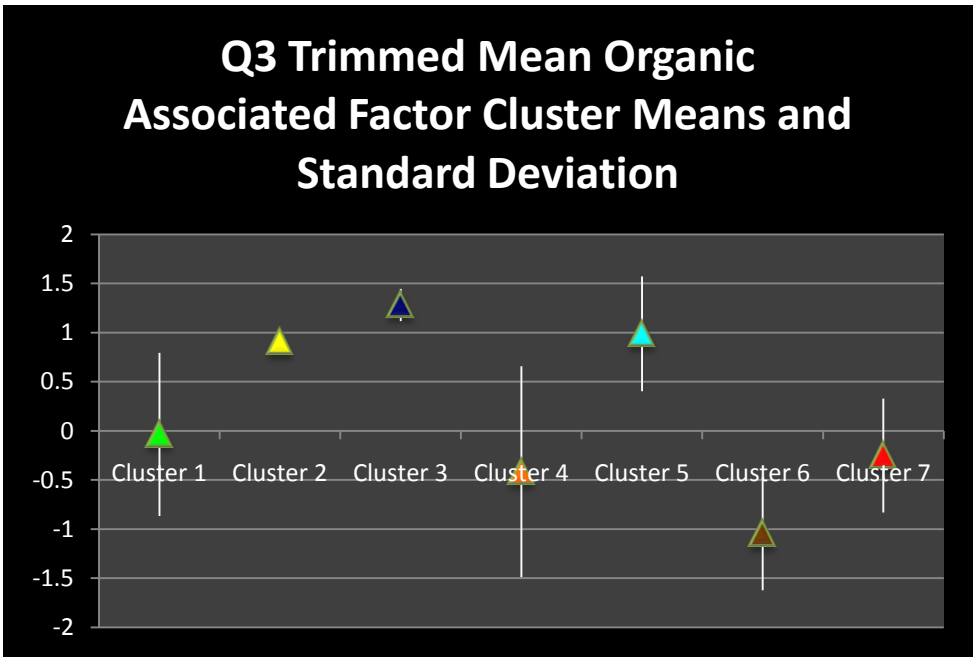
Supplementary Figure 5.40 – Cluster mean comparison for the particle associated factor for the quarter 2 median dataset



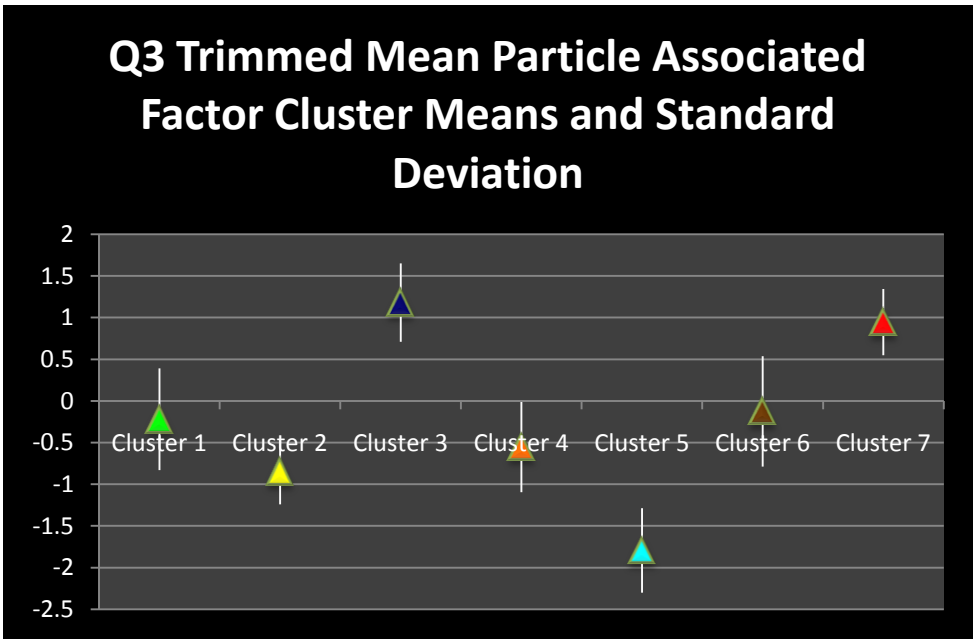
Supplementary Figure 5.41 – Cluster mean comparison for the redox conditions factor for the quarter 2 median dataset



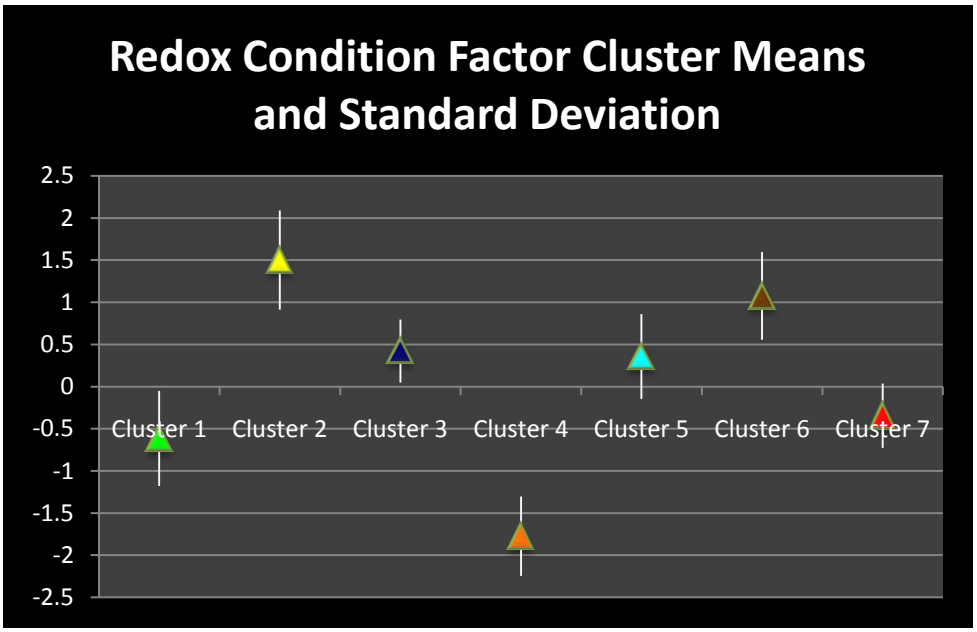
Supplementary Figure 5.42 – Cluster mean comparison for the subsurface flow associated factor for the quarter 2 trimmed mean dataset



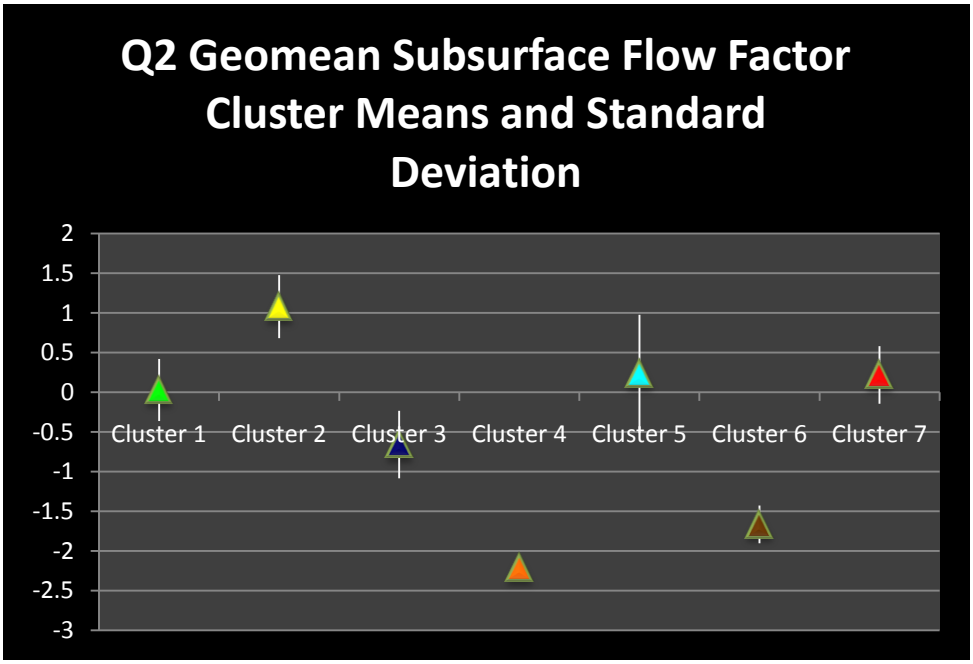
Supplementary Figure 5.43 – Cluster mean comparison for the organics associated factor for the quarter 2 trimmed mean dataset



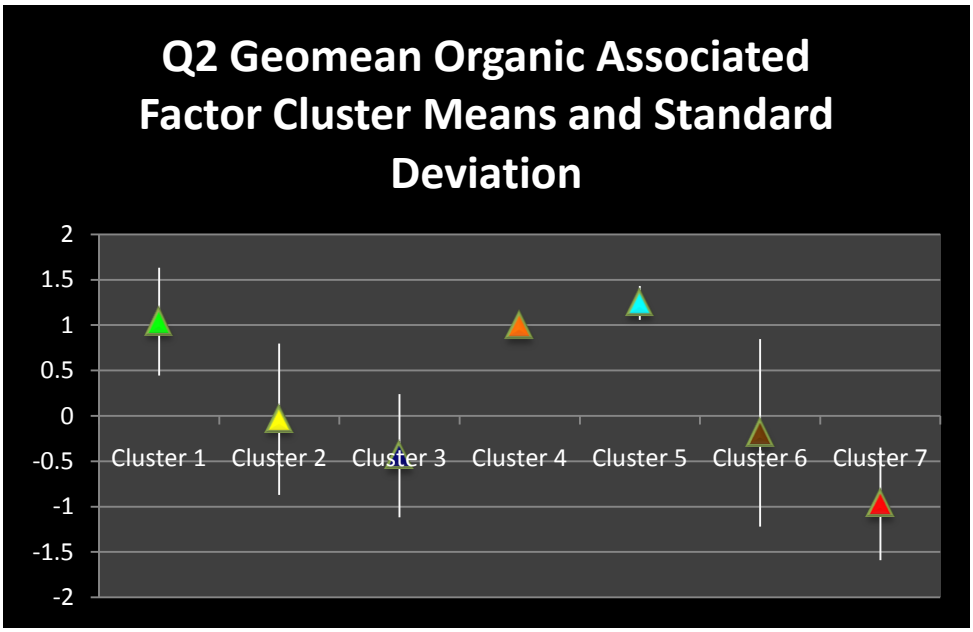
Supplementary Figure 5.44 – Cluster mean comparison for the particle associated factor for the quarter 2 trimmed mean dataset



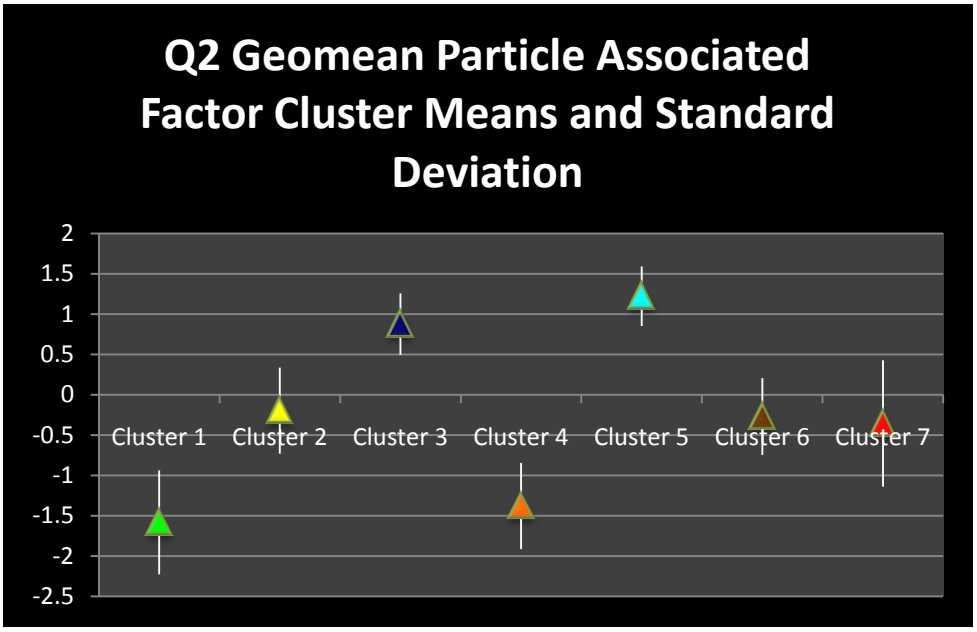
Supplementary Figure 5.45 – Cluster mean comparison for the redox conditions factor for the quarter 2 trimmed mean dataset



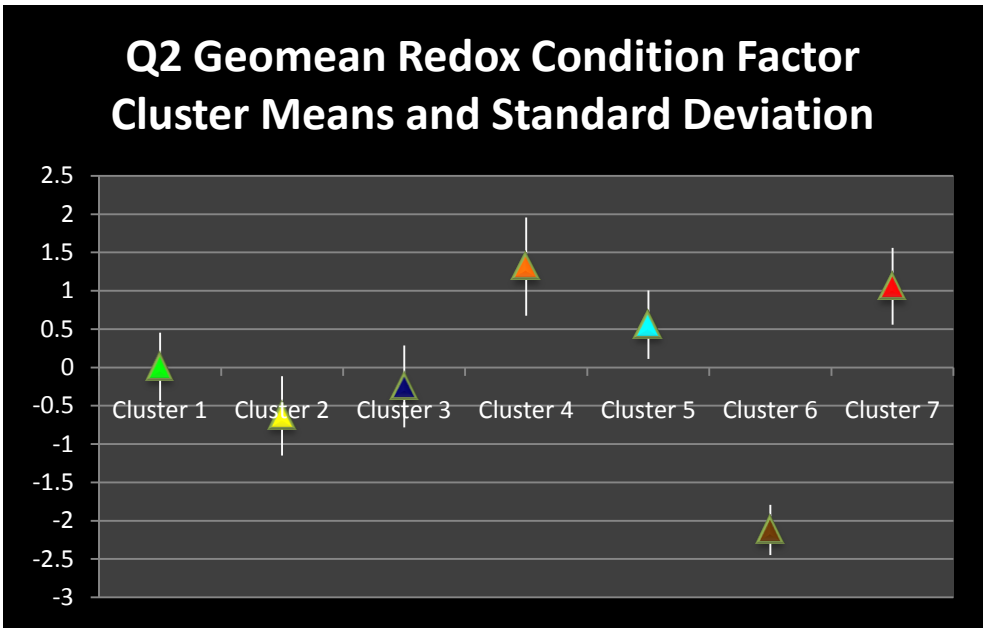
Supplementary Figure 5.46 – Cluster mean comparison for the subsurface flow associated factor for the quarter 2 geometric mean dataset



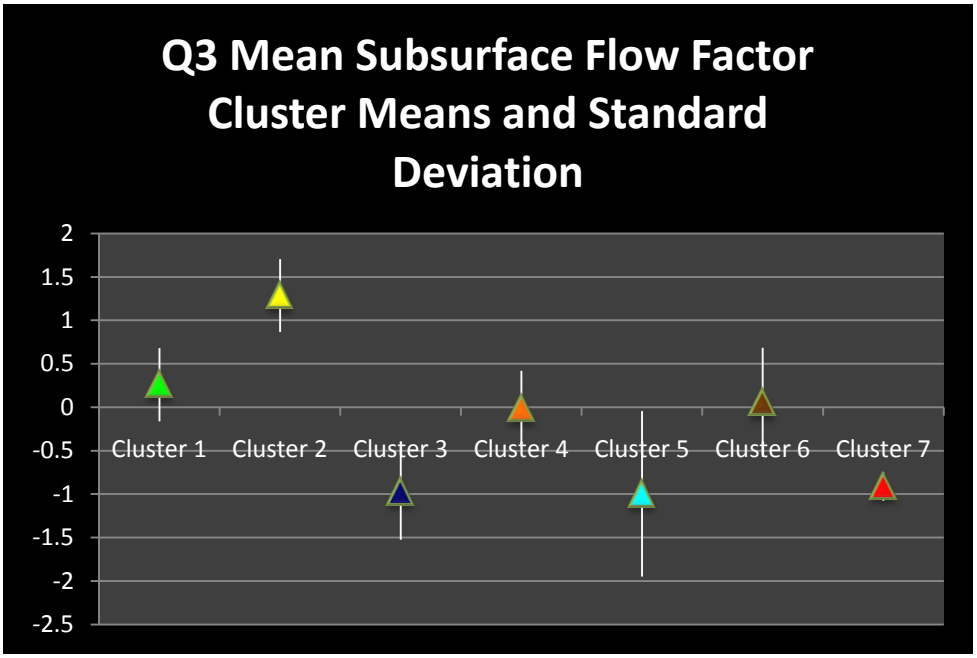
Supplementary Figure 5.47 – Cluster mean comparison for the organics associated factor for the quarter 2 geometric mean dataset



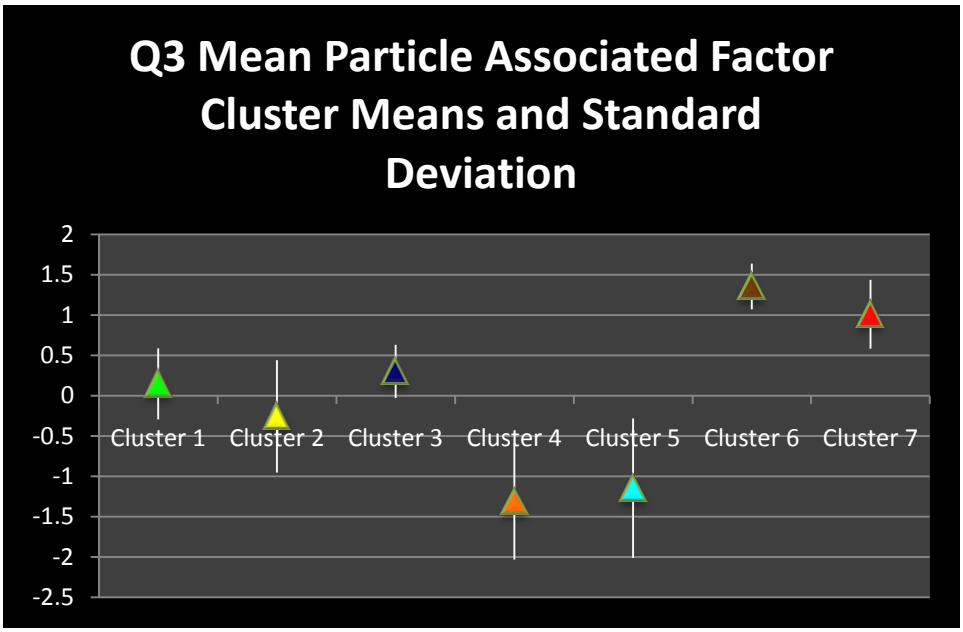
Supplementary Figure 5.48 – Cluster mean comparison for the particle associated factor for the quarter 2 geometric mean dataset



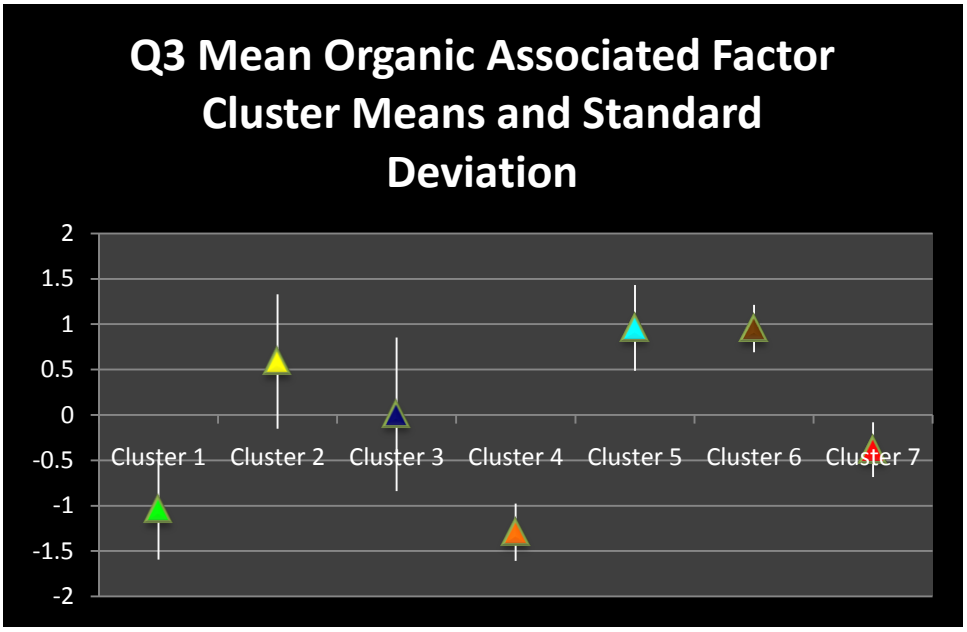
Supplementary Figure 5.49 – Cluster mean comparison for the redox conditions factor for the quarter 2 geometric mean dataset



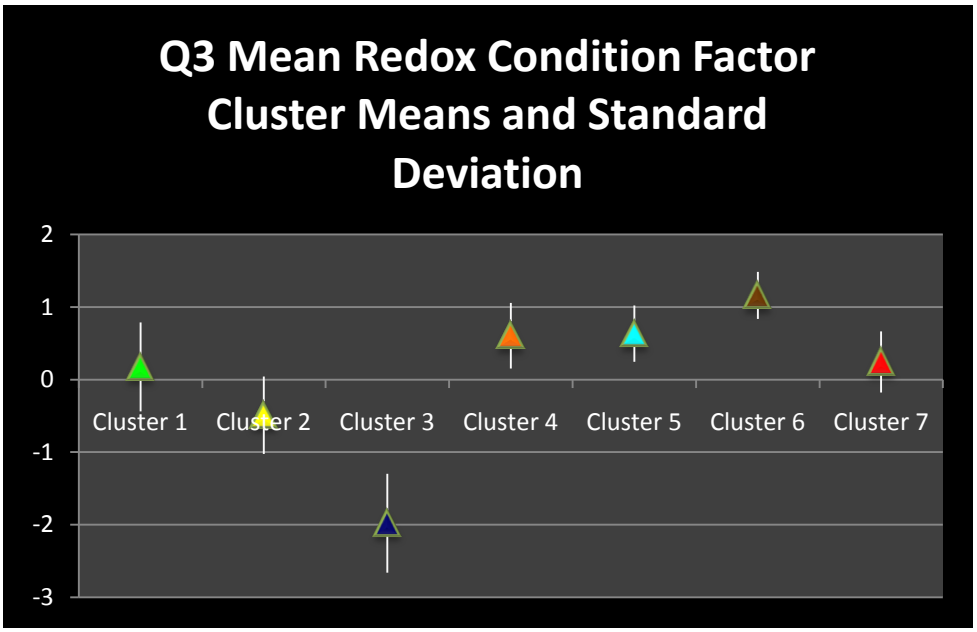
Supplementary Figure 5.50 – Cluster mean comparison for the subsurface flow associated factor for the quarter 3 mean dataset



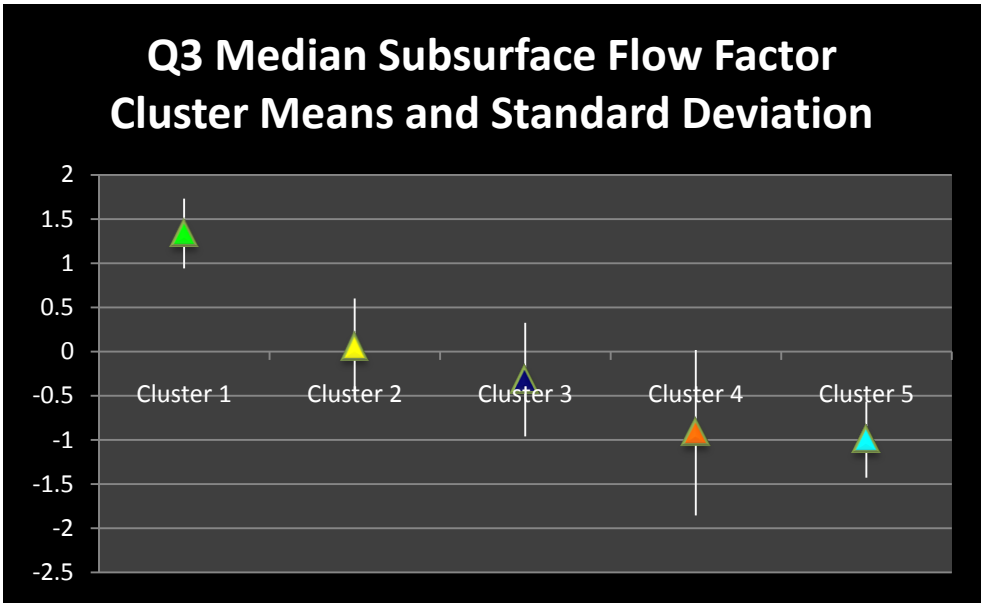
Supplementary Figure 5.51 – Cluster mean comparison for the organics associated factor for the quarter 3 mean dataset



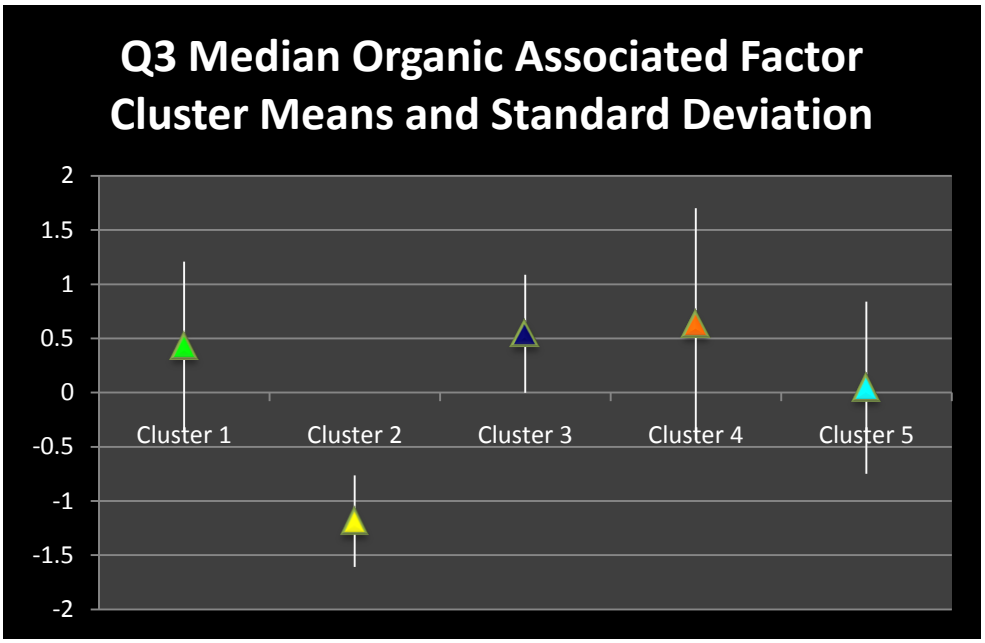
Supplementary Figure 5.52 – Cluster mean comparison for the particle associated factor for the quarter 3 mean dataset



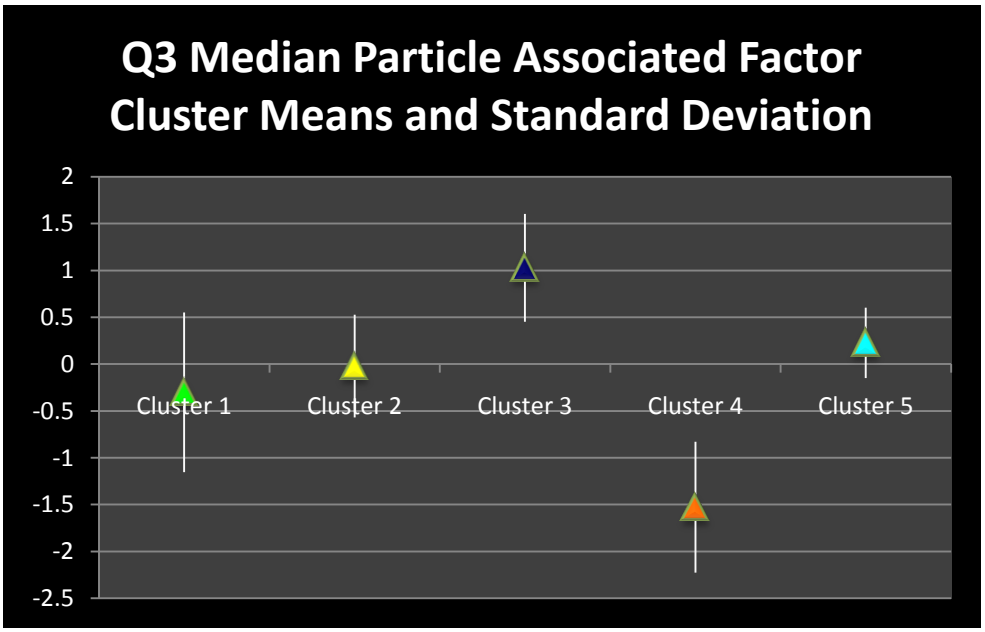
Supplementary Figure 5.53 – Cluster mean comparison for the redox conditions factor for the quarter 3 mean dataset



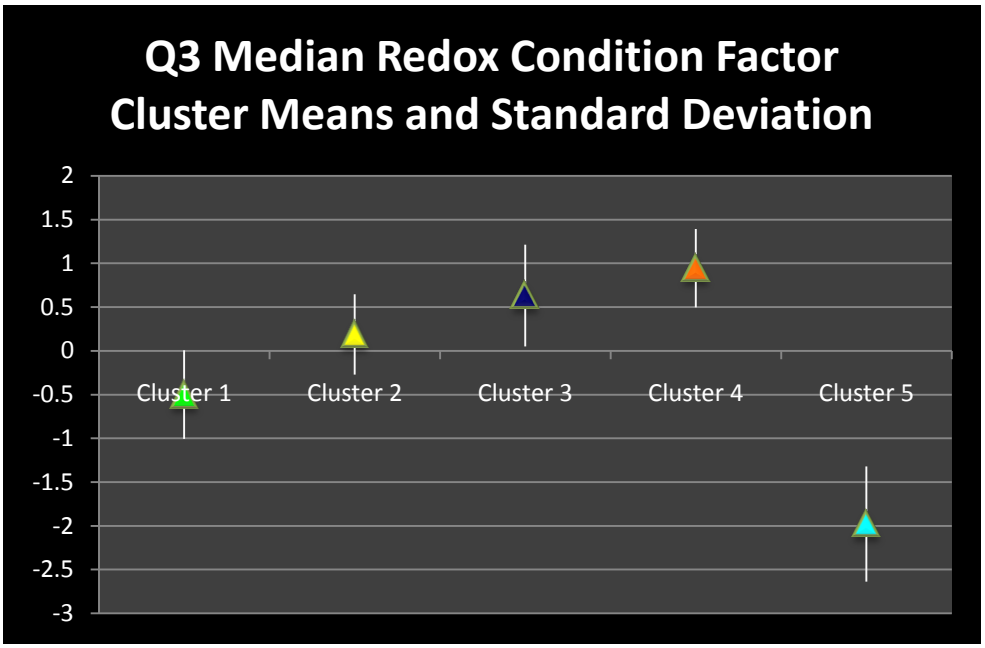
Supplementary Figure 5.54 – Cluster mean comparison for the subsurface flow associated factor for the quarter 3 median dataset



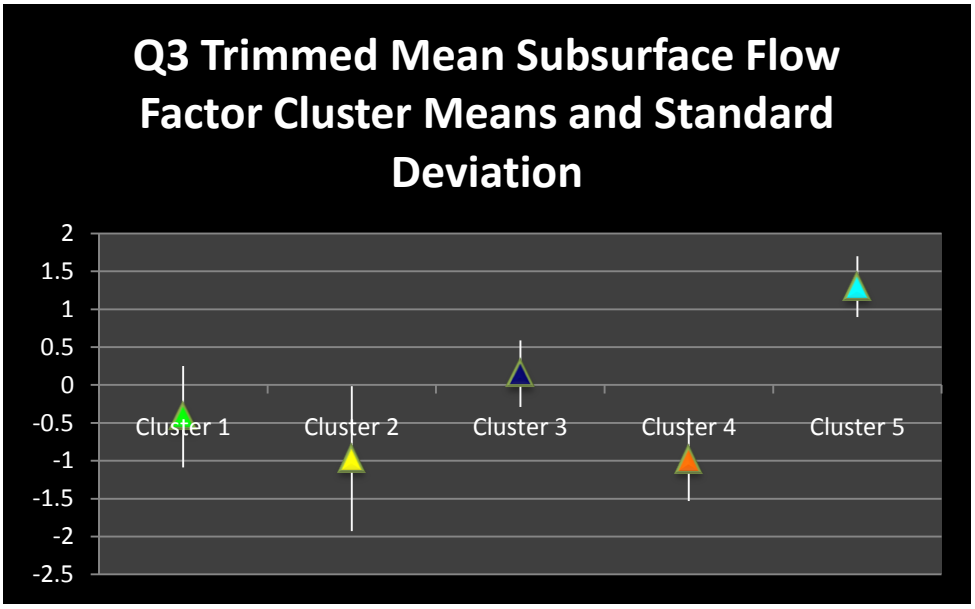
Supplementary Figure 5.55 – Cluster mean comparison for the organics associated factor for the quarter 3 median dataset



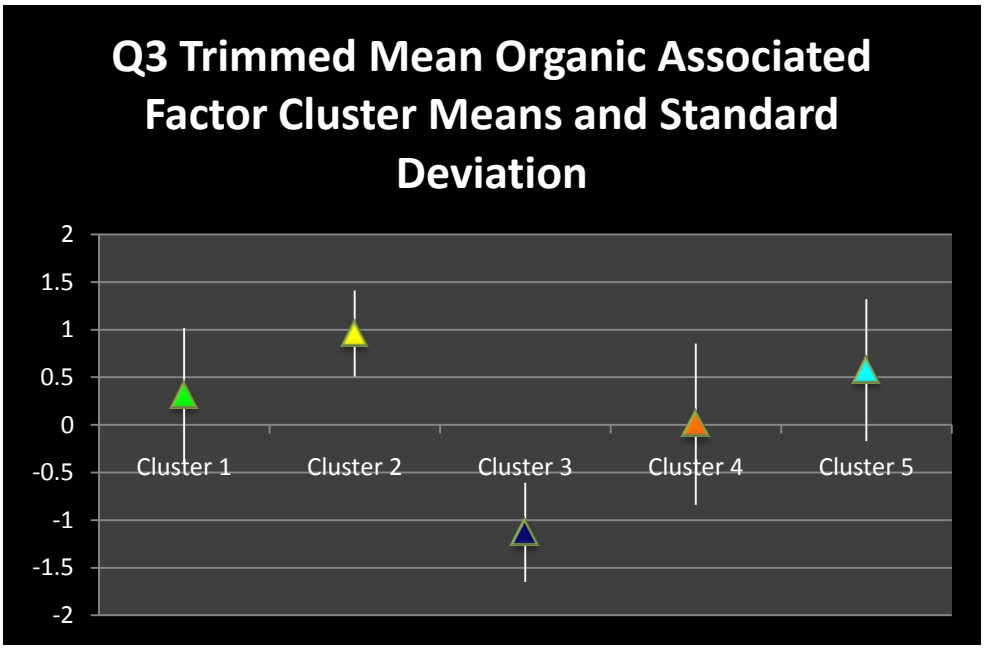
Supplementary Figure 5.56 – Cluster mean comparison for the particle associated factor for the quarter 3 median dataset



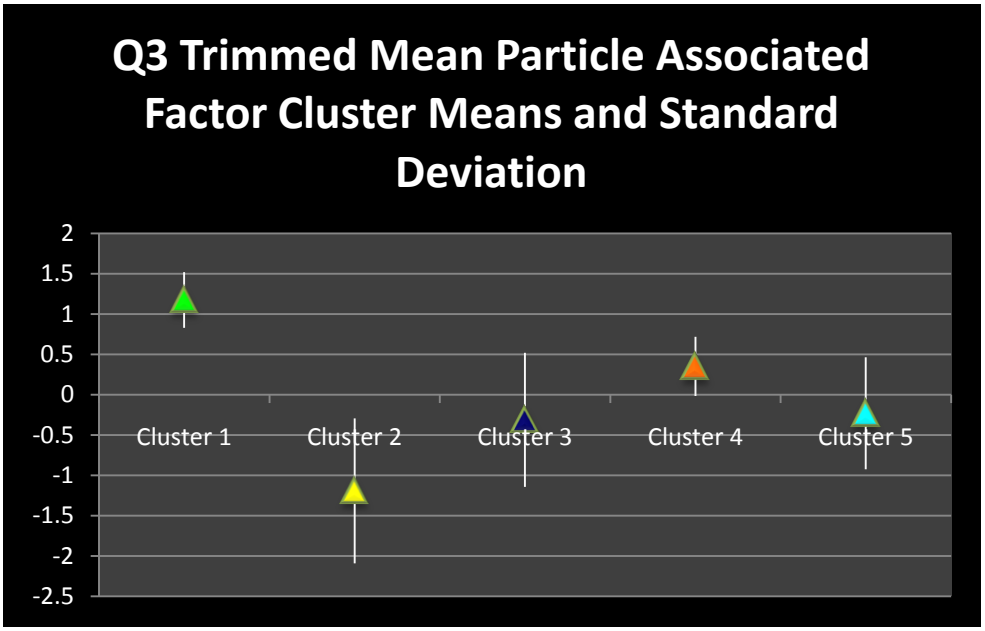
Supplementary Figure 5.57 – Cluster mean comparison for the redox conditions factor for the quarter 3 median dataset



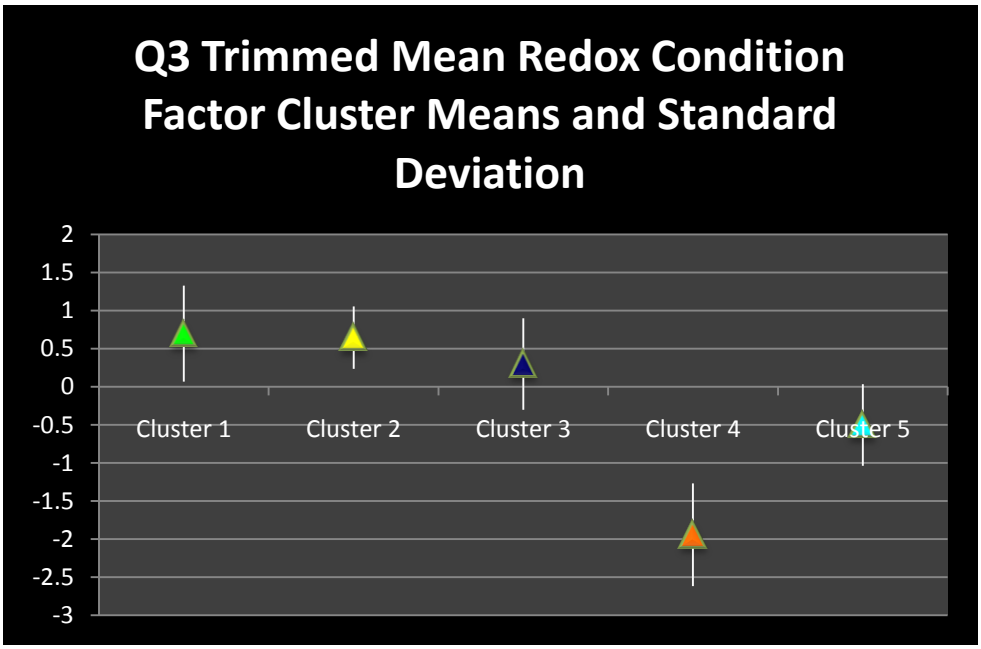
Supplementary Figure 5.58 – Cluster mean comparison for the subsurface flow associated factor for the quarter 3 trimmed mean dataset



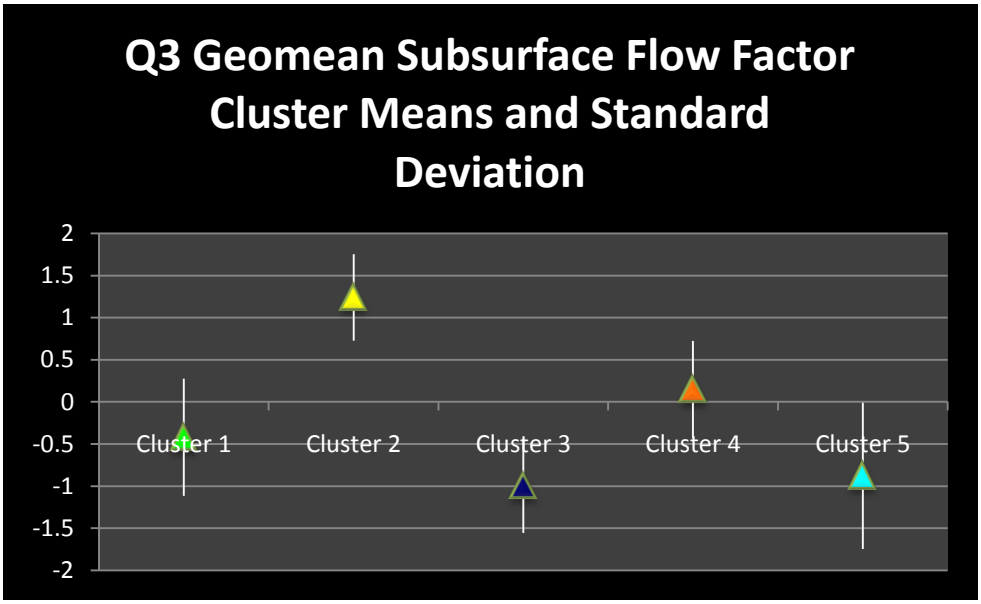
Supplementary Figure 5.59 – Cluster mean comparison for the organics associated factor for the quarter 3 trimmed mean dataset



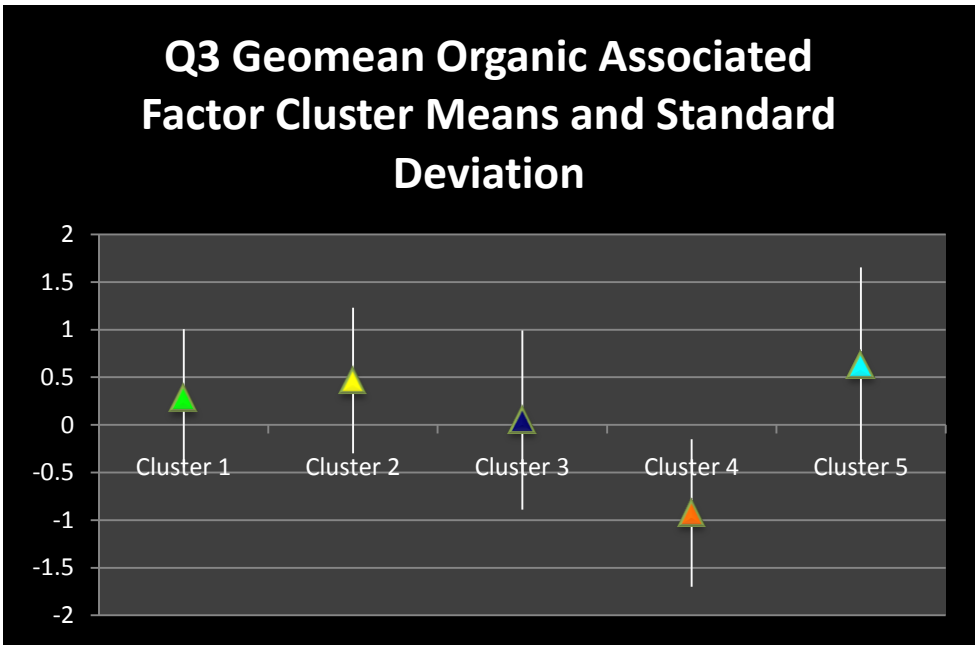
Supplementary Figure 5.60 – Cluster mean comparison for the particle associated factor for the quarter 3 trimmed mean dataset



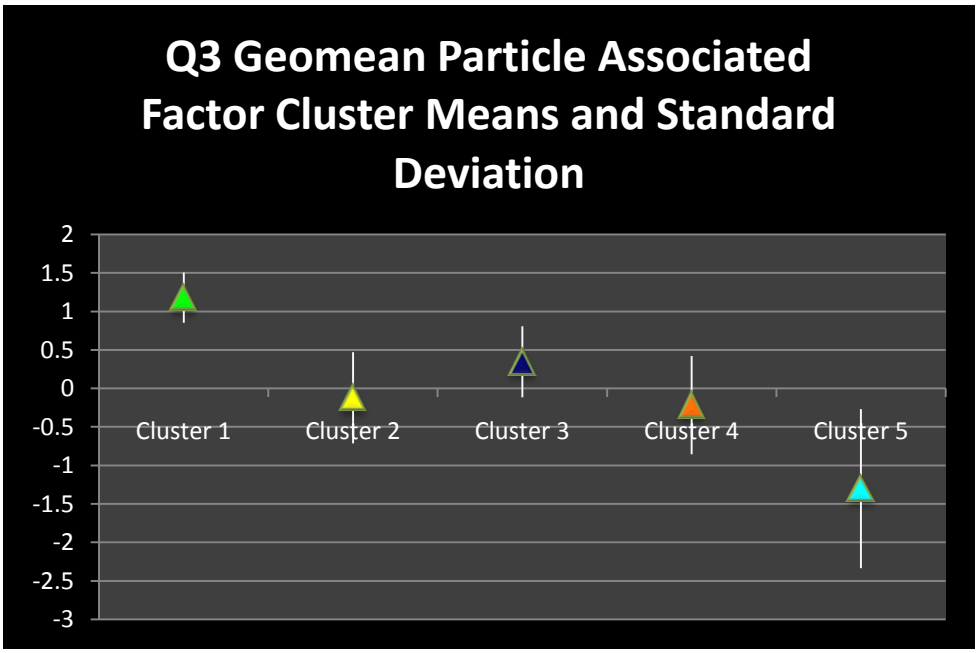
Supplementary Figure 5.61 – Cluster mean comparison for the redox conditions factor for the quarter 3 trimmed mean dataset



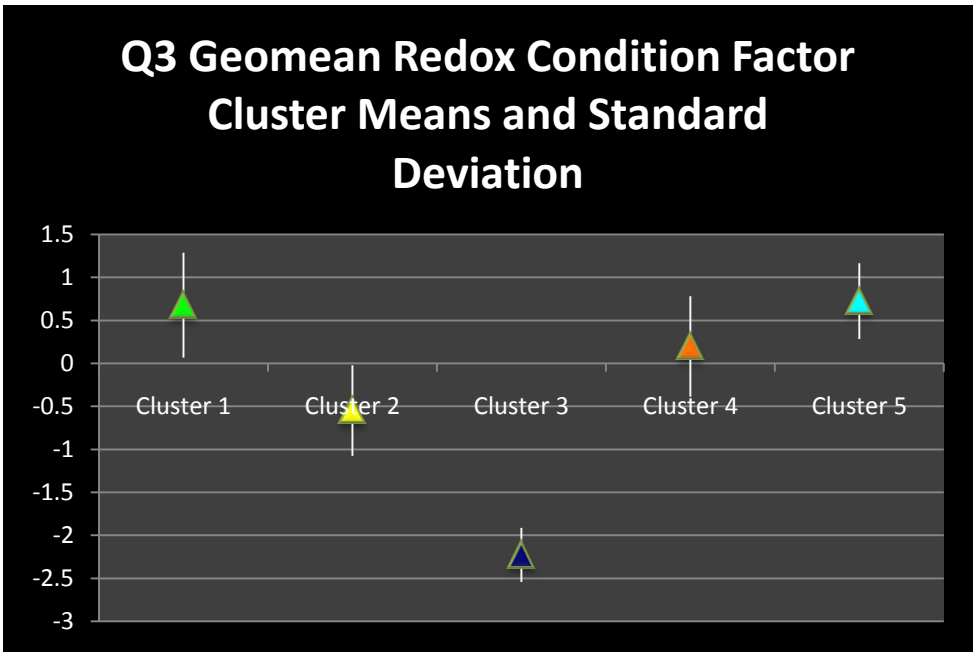
Supplementary Figure 5.62 – Cluster mean comparison for the subsurface flow associated factor for the quarter 3 geometric mean dataset



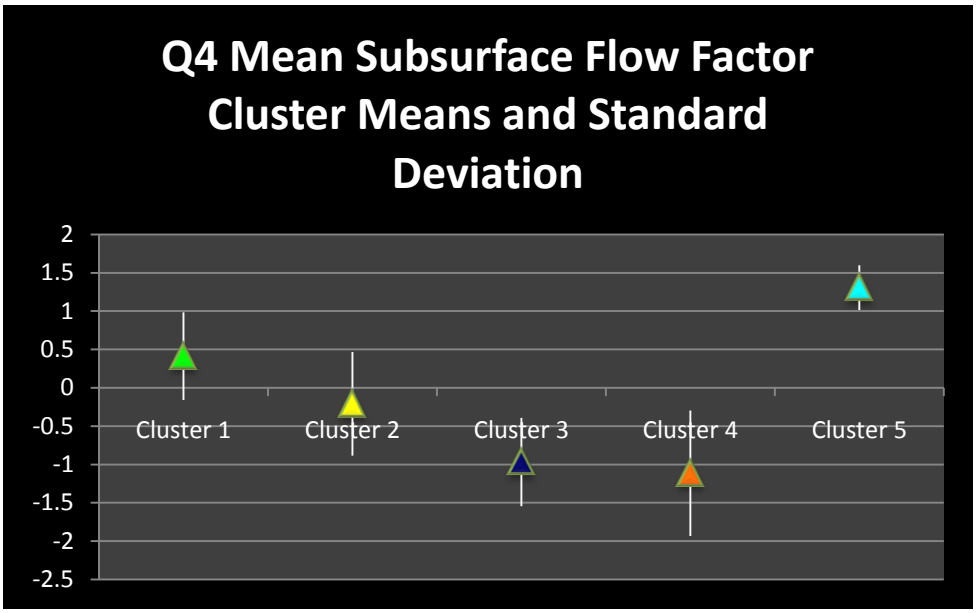
Supplementary Figure 5.63 – Cluster mean comparison for the organics associated factor for the quarter 3 geometric mean dataset



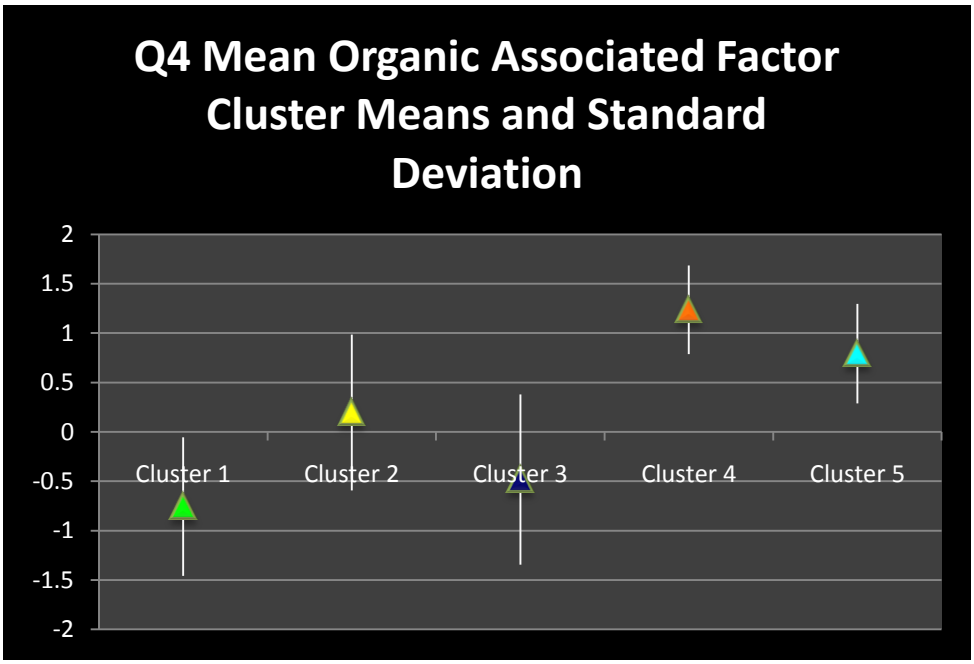
Supplementary Figure 5.64 – Cluster mean comparison for the particle associated factor for the quarter 3 geometric mean dataset



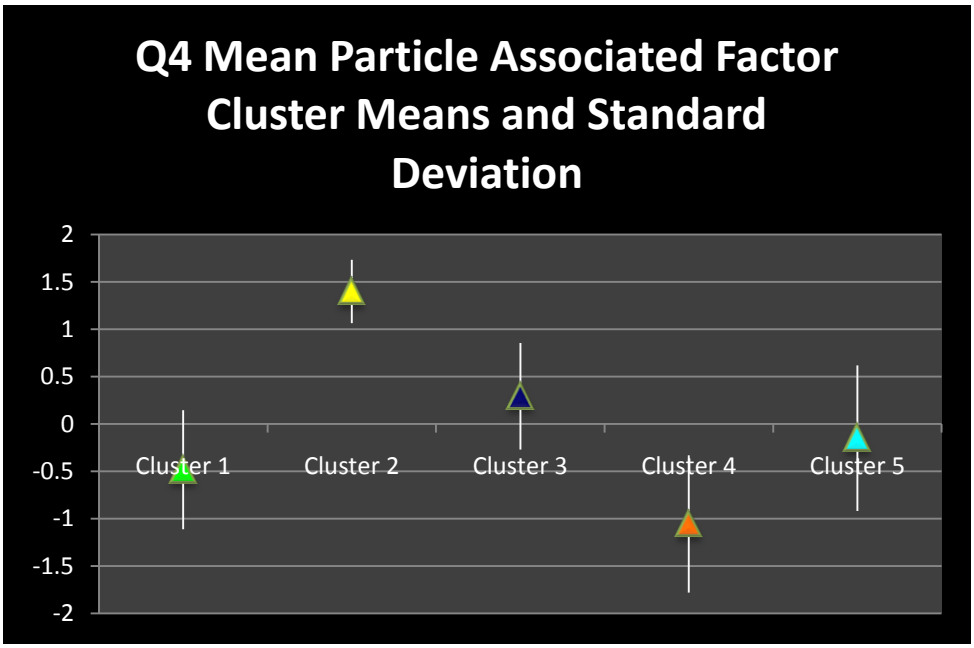
Supplementary Figure 5.65 – Cluster mean comparison for the redox conditions factor for the quarter 3 geometric mean dataset



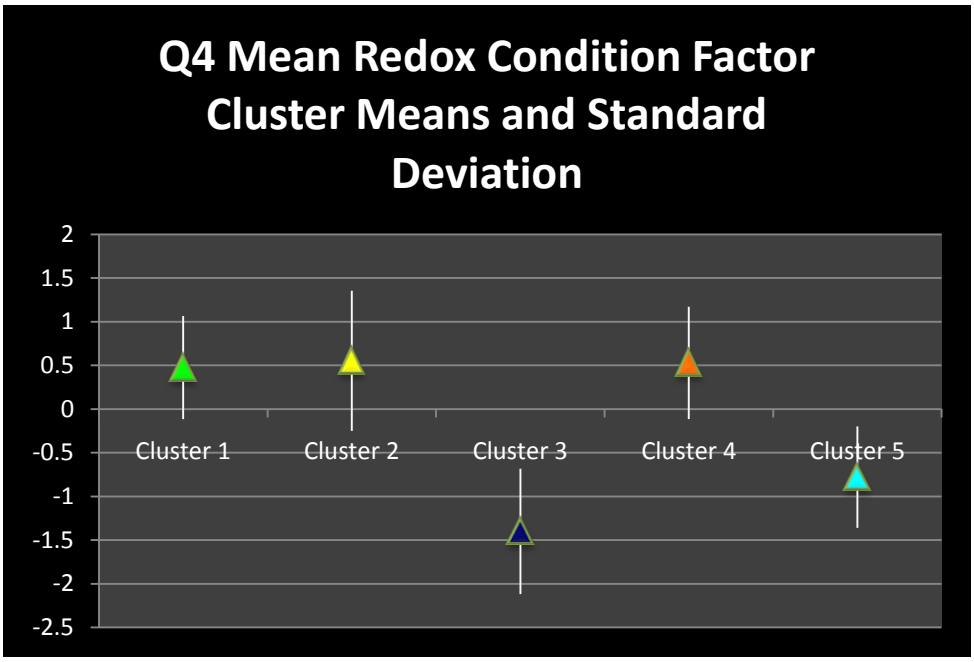
Supplementary Figure 5.66 – Cluster mean comparison for the subsurface flow associated factor for the quarter 4 mean dataset



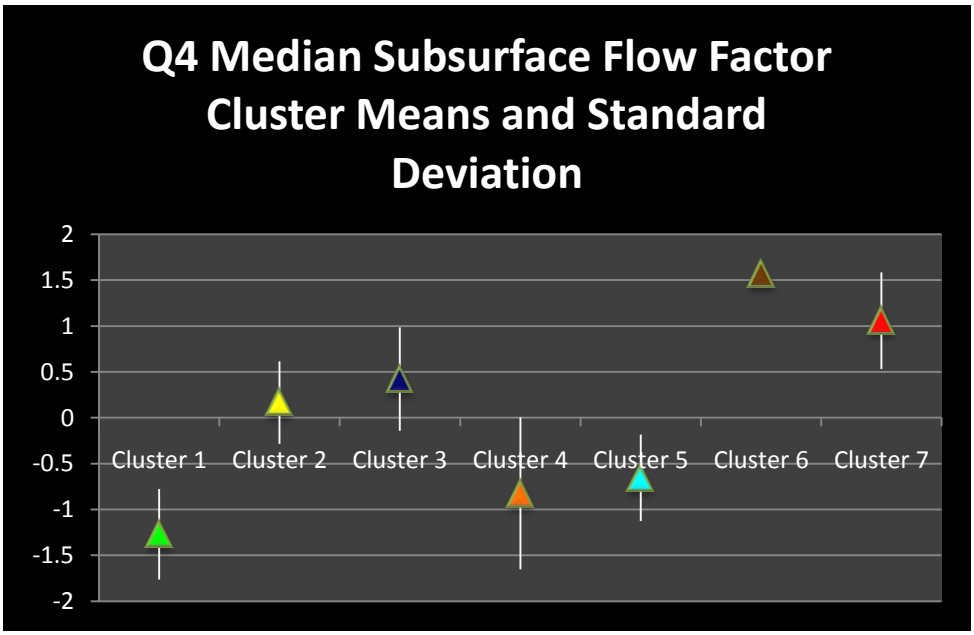
Supplementary Figure 5.67 – Cluster mean comparison for the organics associated factor for the quarter 4 mean dataset



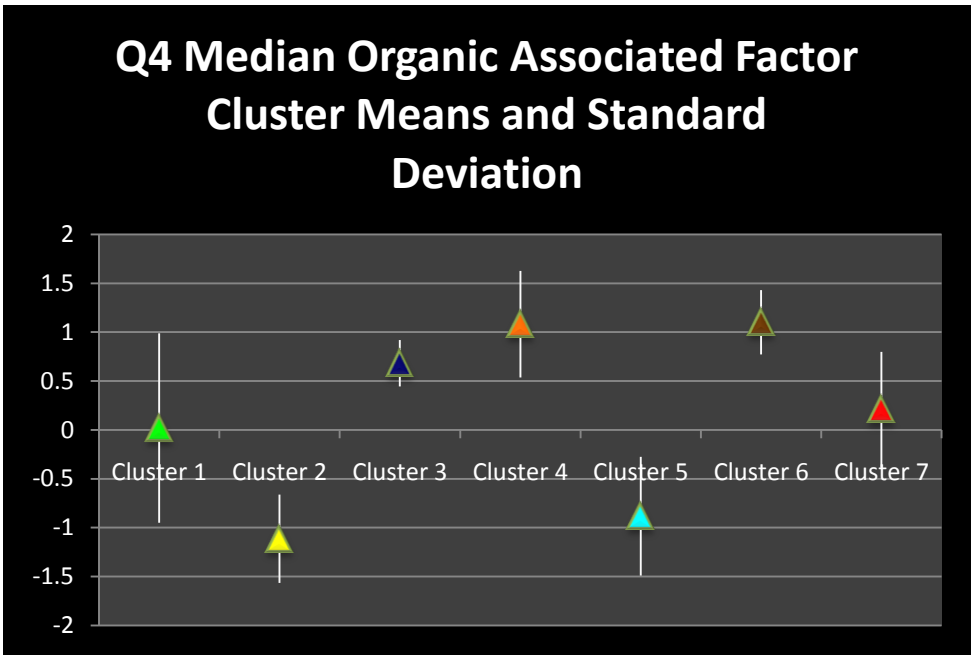
Supplementary Figure 5.68 – Cluster mean comparison for the particle associated factor for the quarter 4 mean dataset



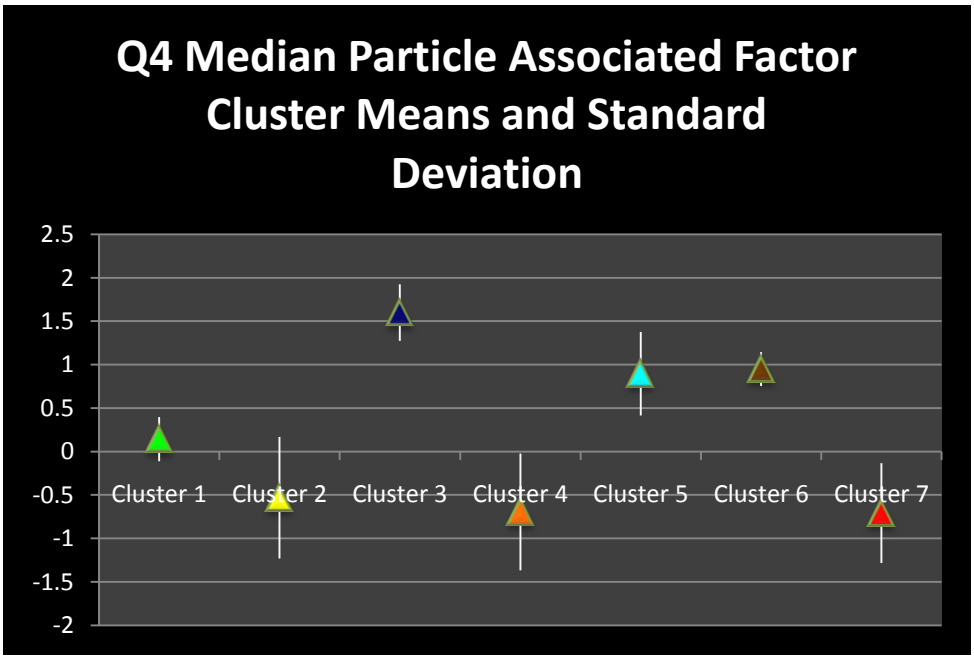
Supplementary Figure 5.69 – Cluster mean comparison for the redox conditions factor for the quarter 4 mean dataset



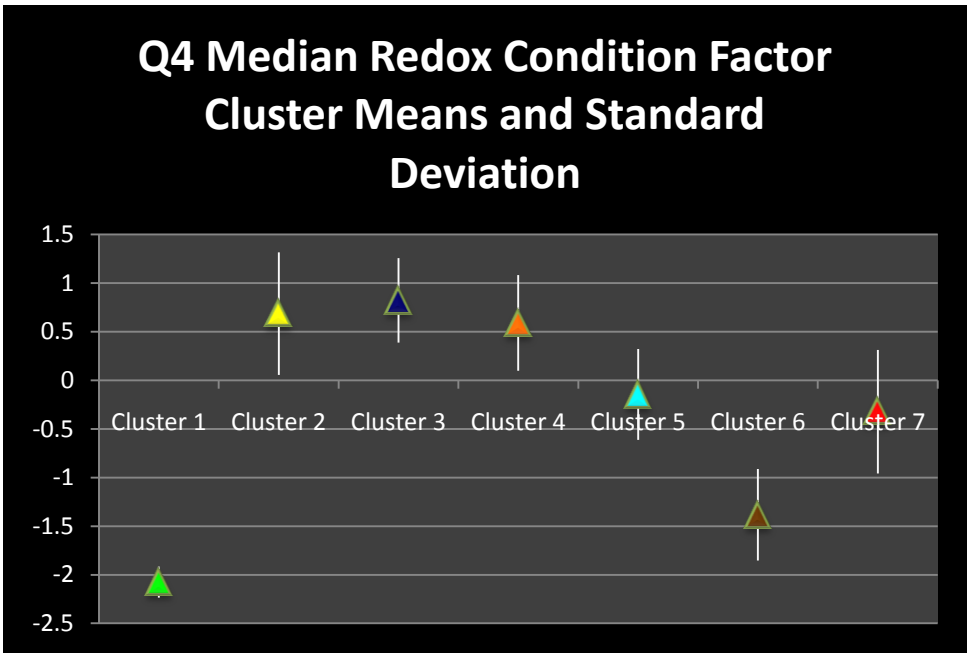
Supplementary Figure 5.70 – Cluster mean comparison for the subsurface flow associated factor for the quarter 4 median dataset



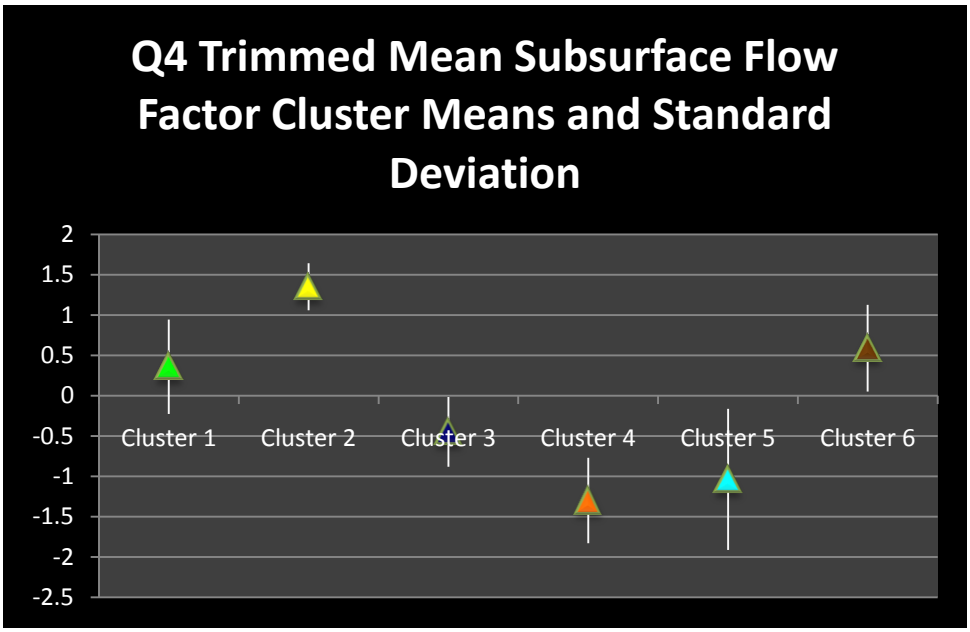
Supplementary Figure 5.71 – Cluster mean comparison for the organics associated factor for the quarter 4 median dataset



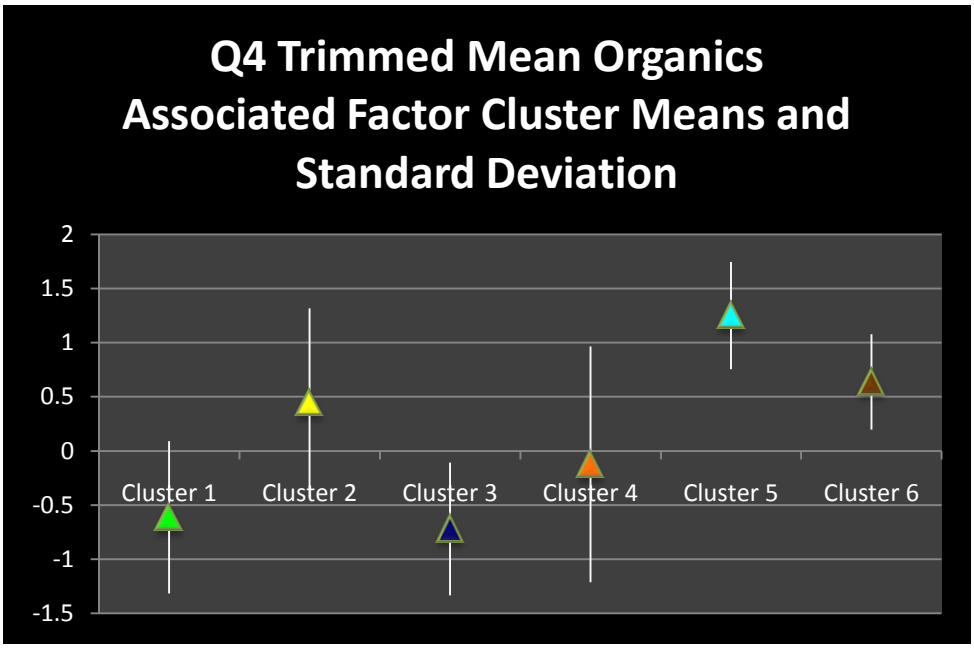
Supplementary Figure 5.72 – Cluster mean comparison for the particle associated factor for the quarter 4 median dataset



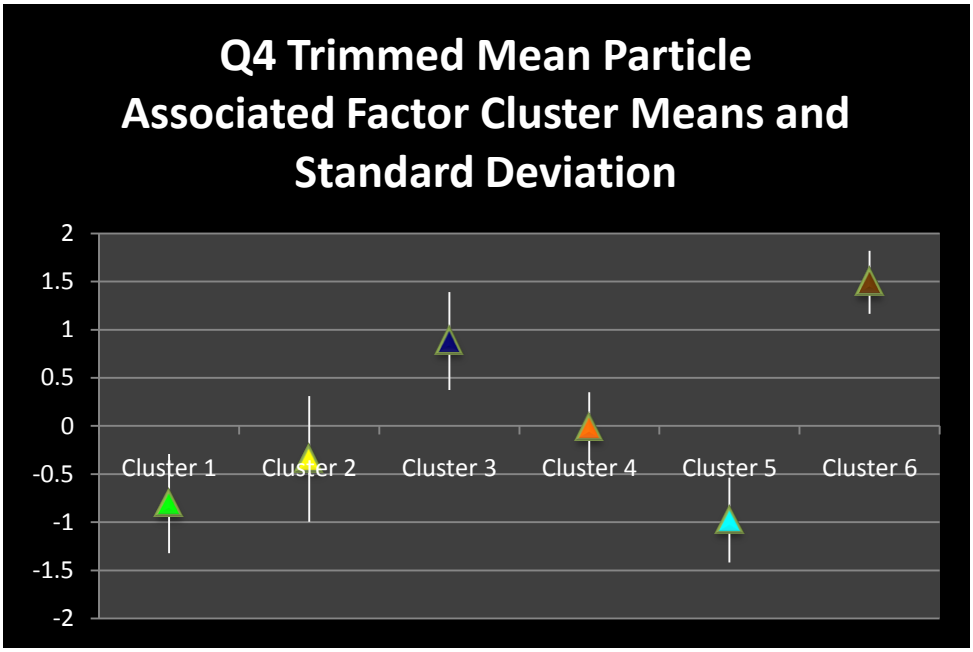
Supplementary Figure 5.73 – Cluster mean comparison for the redox conditions factor for the quarter 4 median dataset



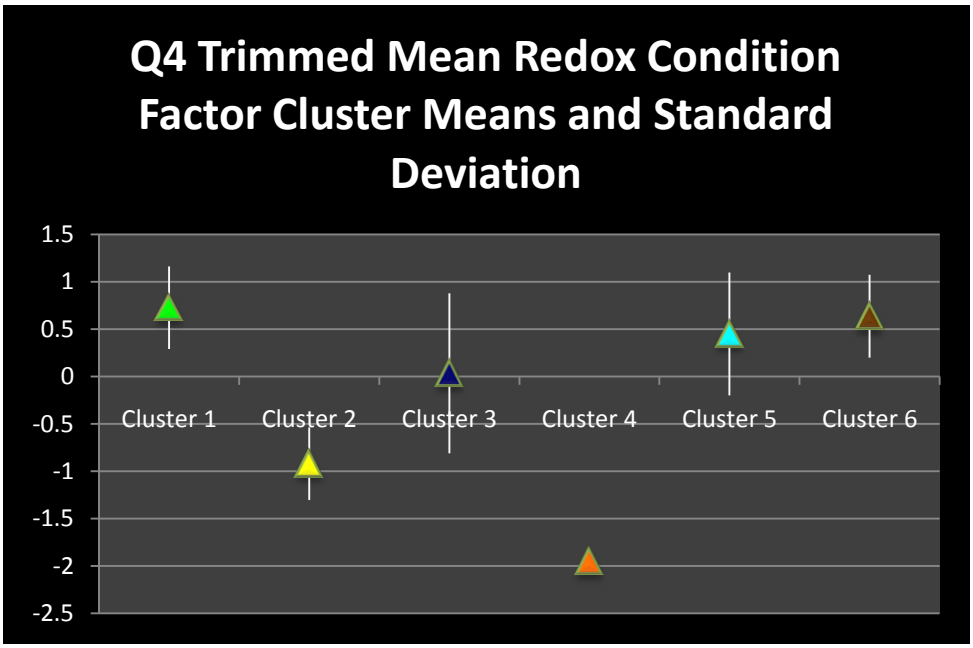
Supplementary Figure 5.74 – Cluster mean comparison for the subsurface flow associated factor for the quarter 4 trimmed mean dataset



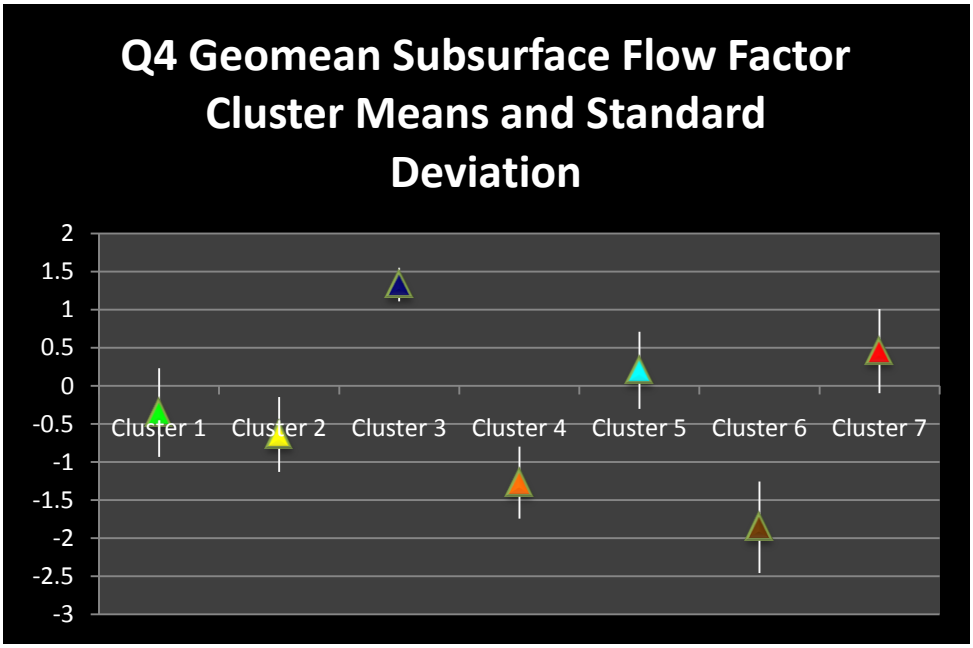
Supplementary Figure 5.75 – Cluster mean comparison for the organics associated factor for the quarter 4 trimmed mean dataset



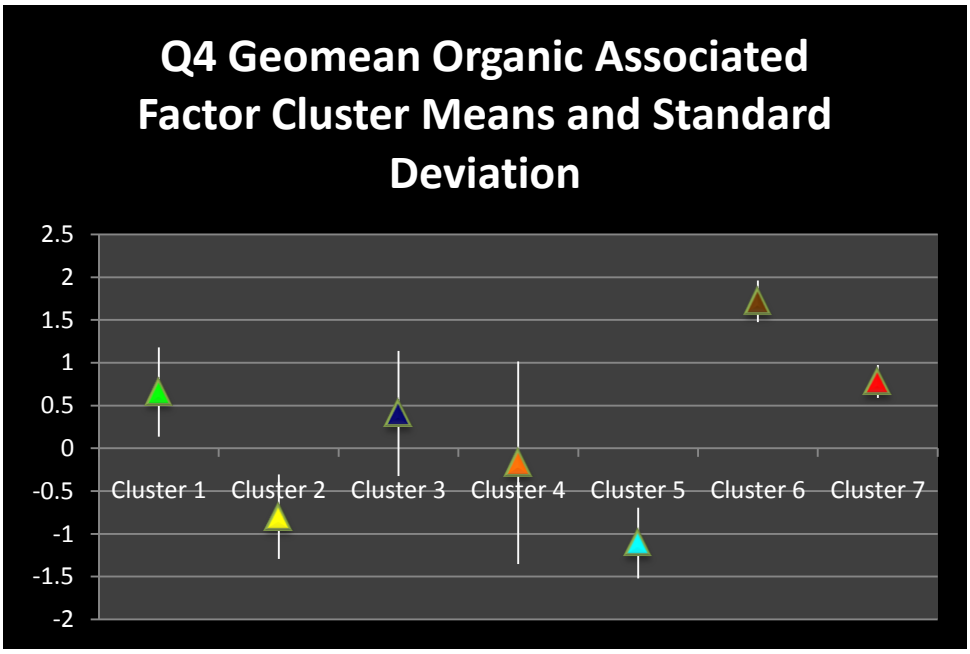
Supplementary Figure 5.76 – Cluster mean comparison for the particle associated factor for the quarter 4 trimmed mean dataset



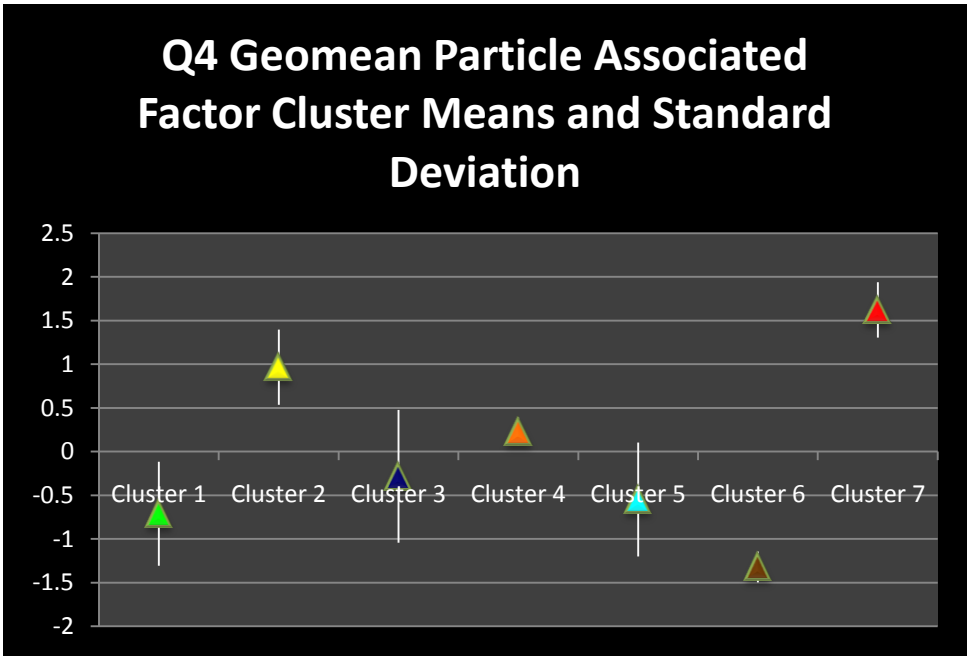
Supplementary Figure 5.77 – Cluster mean comparison for the redox conditions factor for the quarter 4 trimmed mean dataset



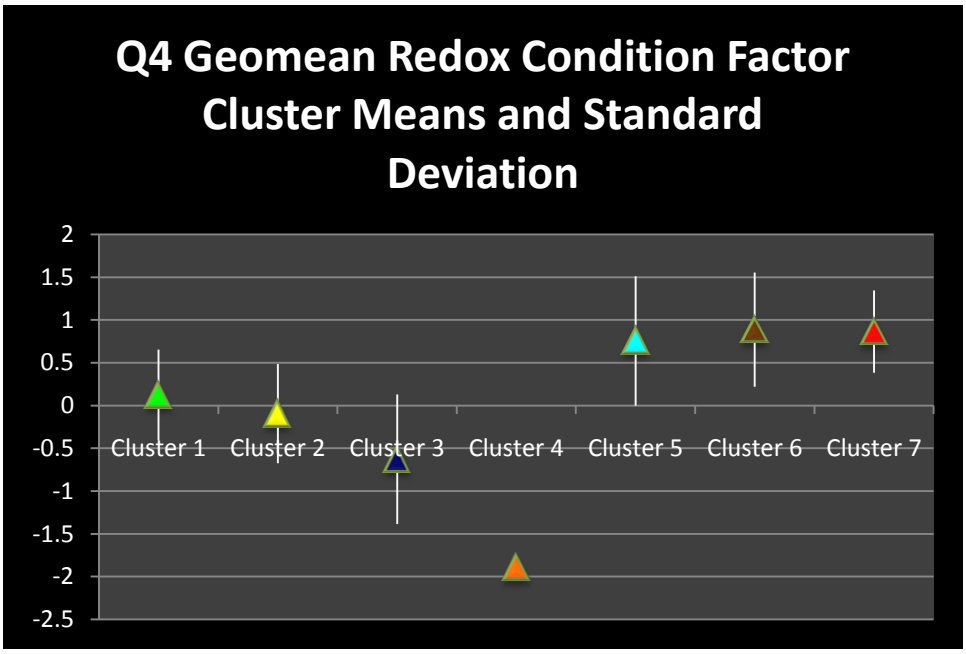
Supplementary Figure 5.78 – Cluster mean comparison for the subsurface flow associated factor for the quarter 4 geometric mean dataset



Supplementary Figure 5.79 – Cluster mean comparison for the organics associated factor for the quarter 4 geometric mean dataset

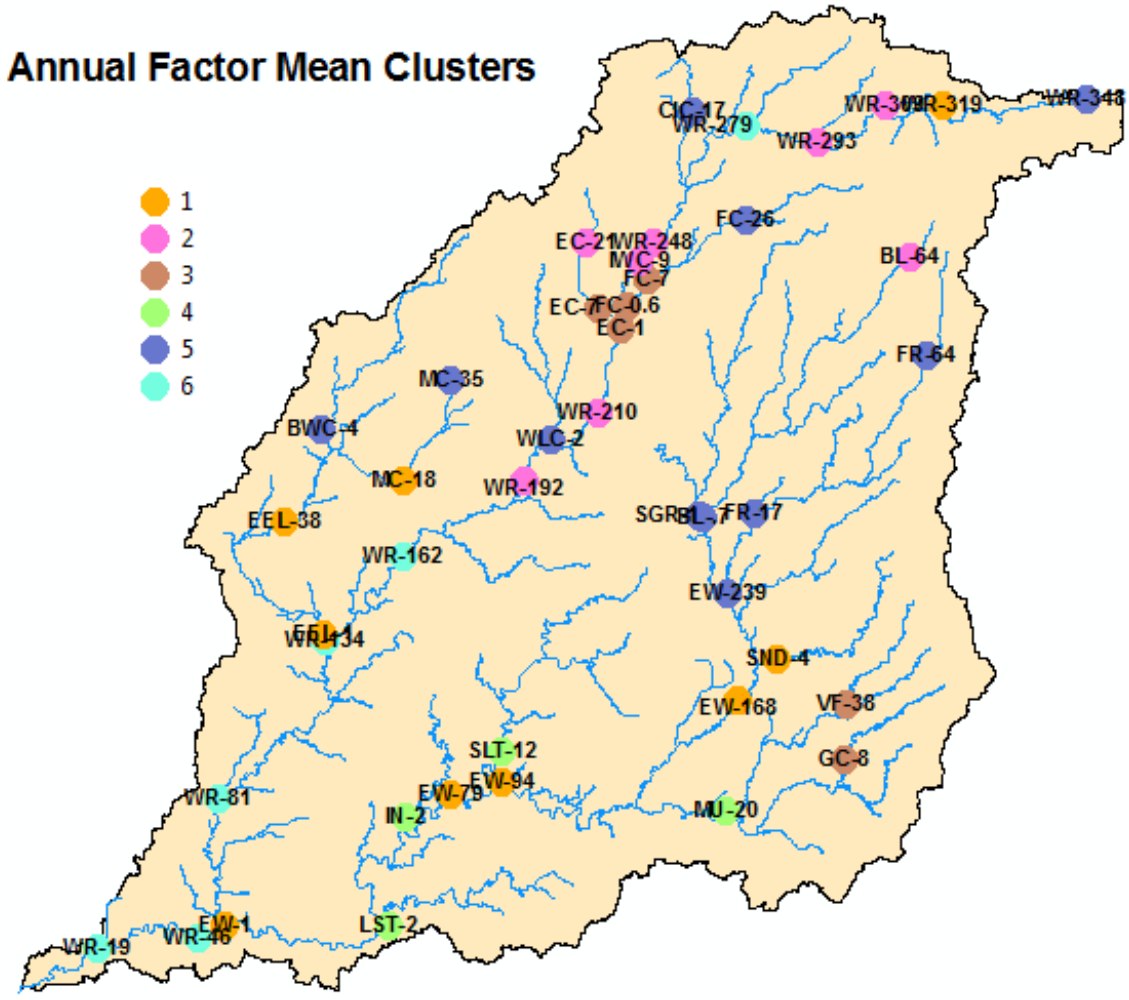


Supplementary Figure 5.80 – Cluster mean comparison for the particle associated factor for the quarter 4 geometric mean dataset

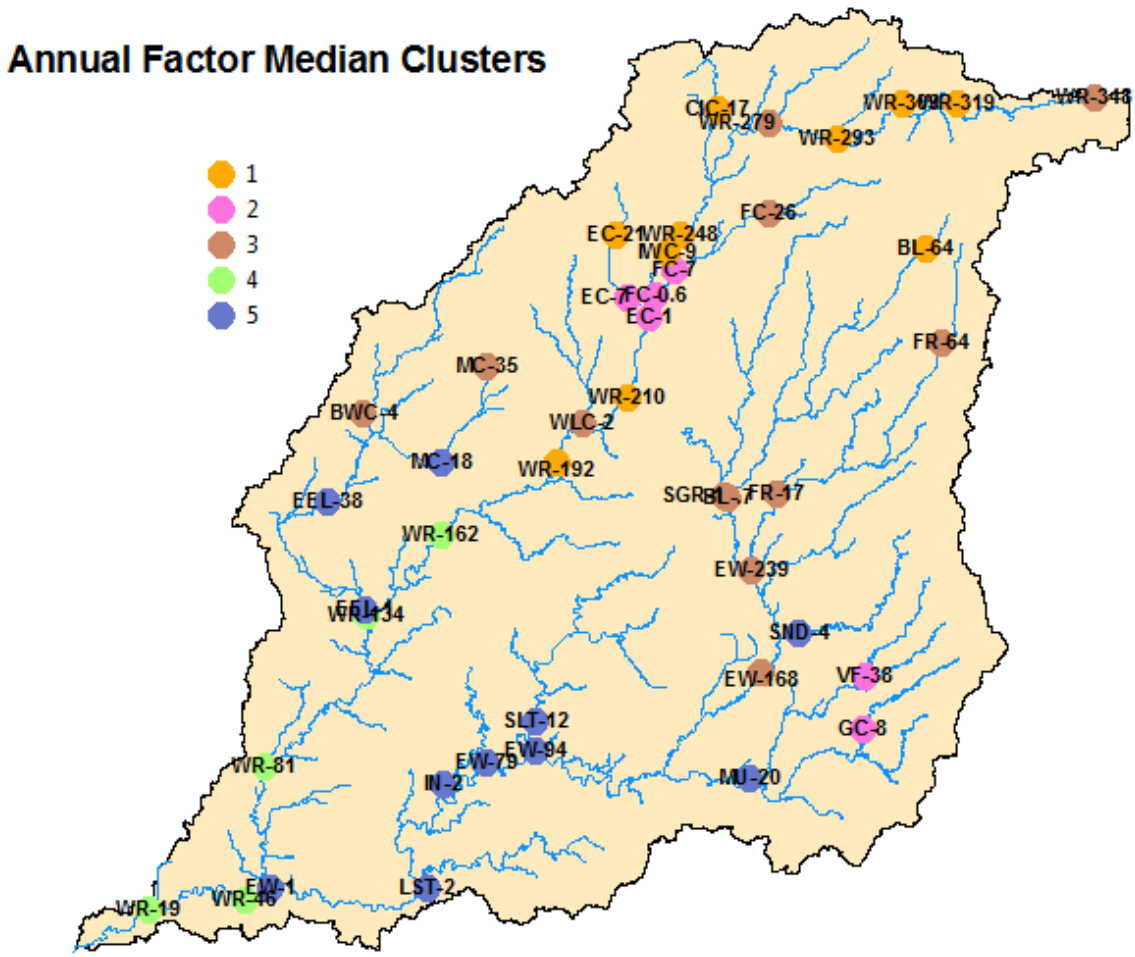


Supplementary Figure 5.81 – Cluster mean comparison for the redox conditions factor for the quarter 4 geometric mean dataset

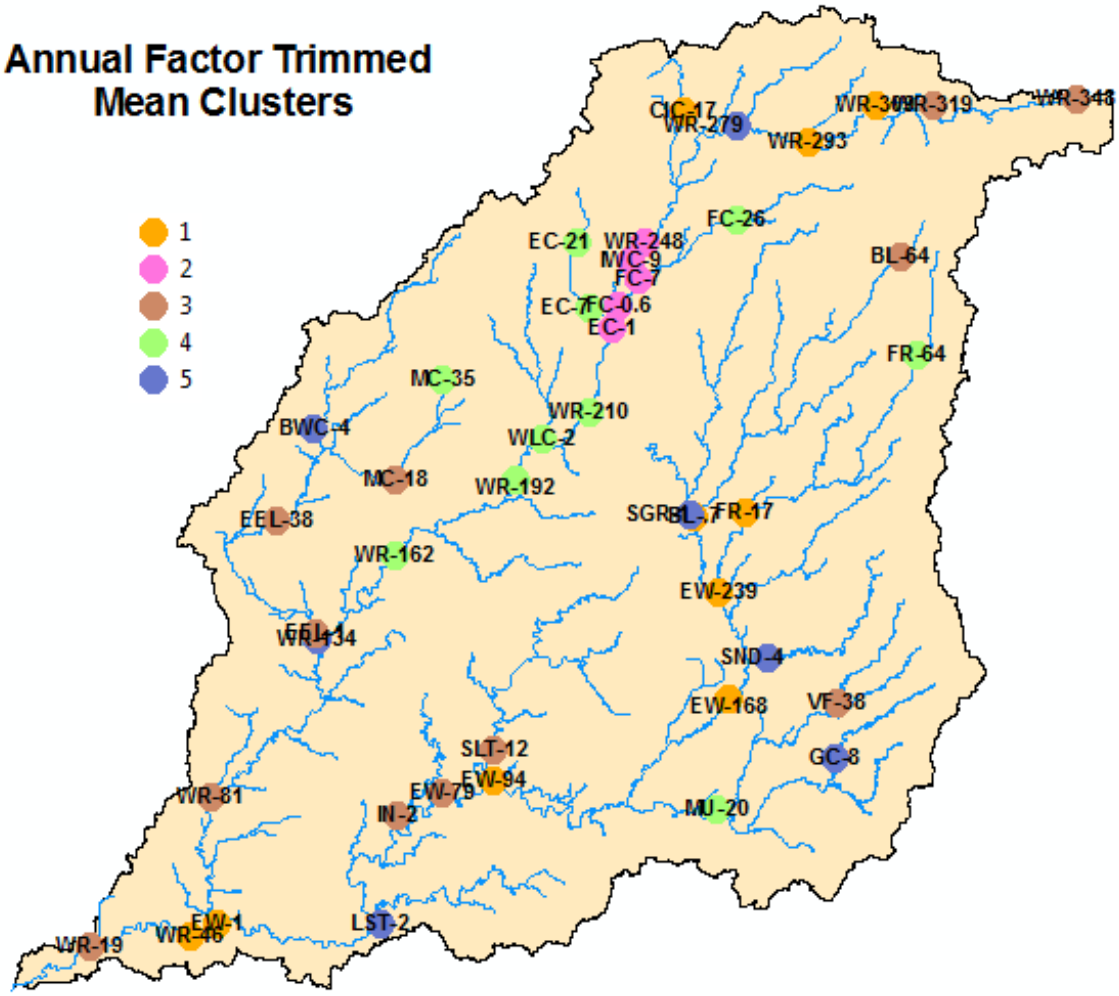
Spatial Distribution of Clusters



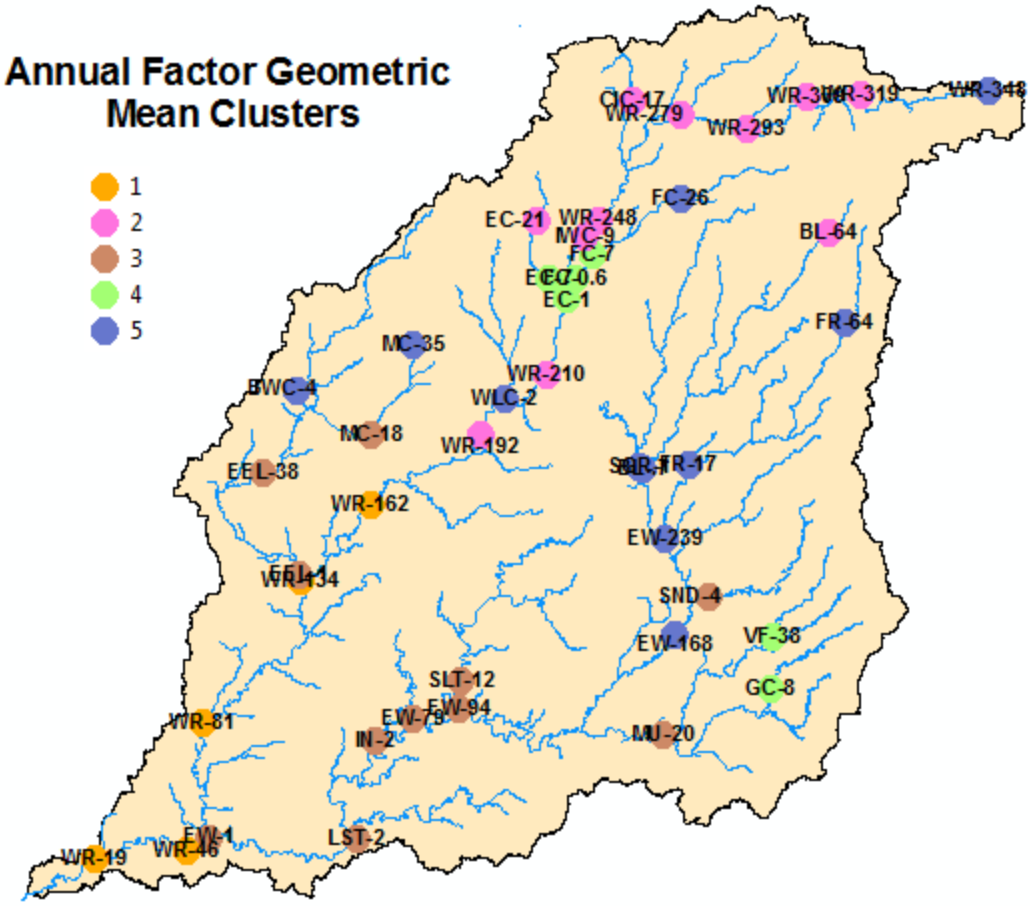
Supplementary Figure 6.1 – Spatially distributed clustering for the Annual Mean factors (EEL-1 belongs to cluster 1 and WR-134 belongs to cluster 6; SGR-1 and BL-.7 both belong to cluster 5)



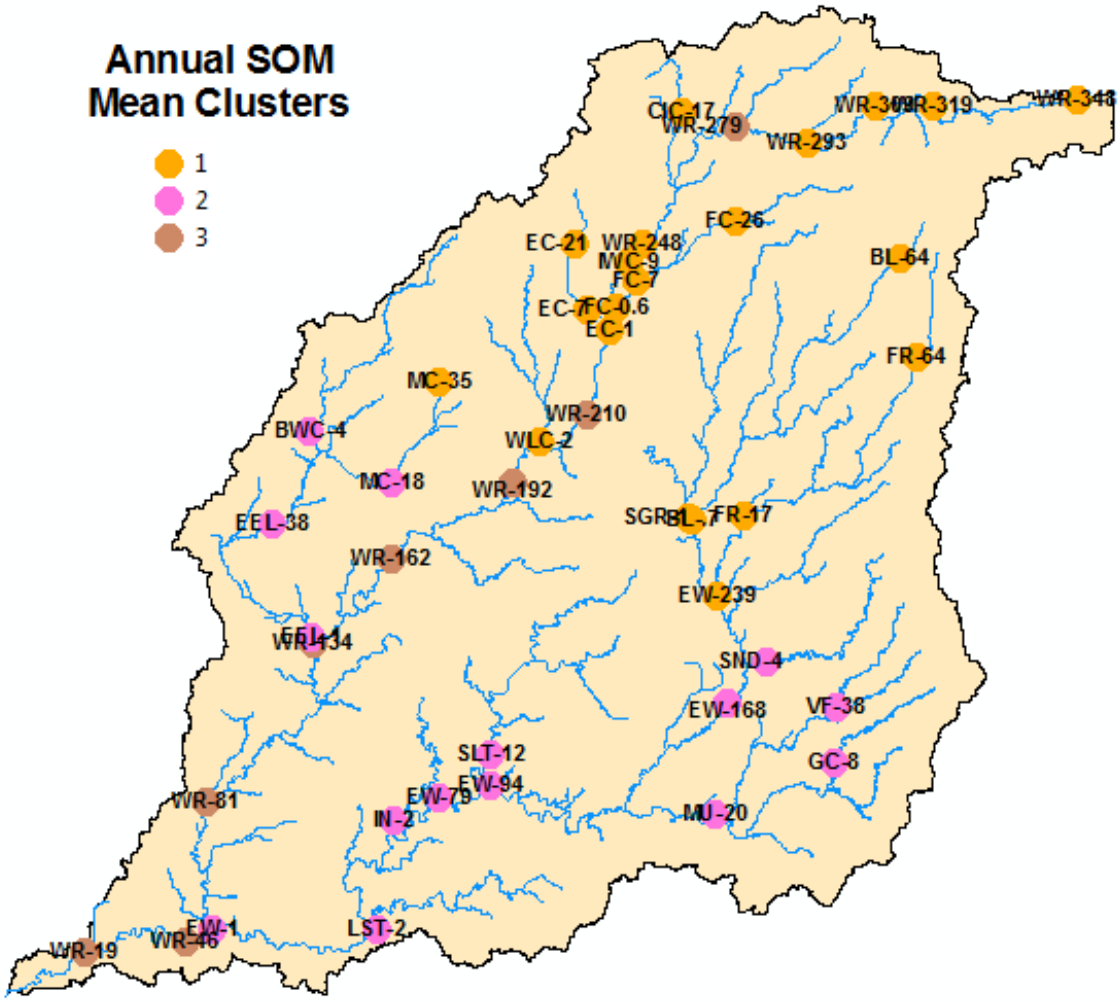
Supplementary Figure 6.2 – Spatially distributed clustering for the Annual Median factors (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 4; SGR-1 and BL-.7 belong to cluster 3)



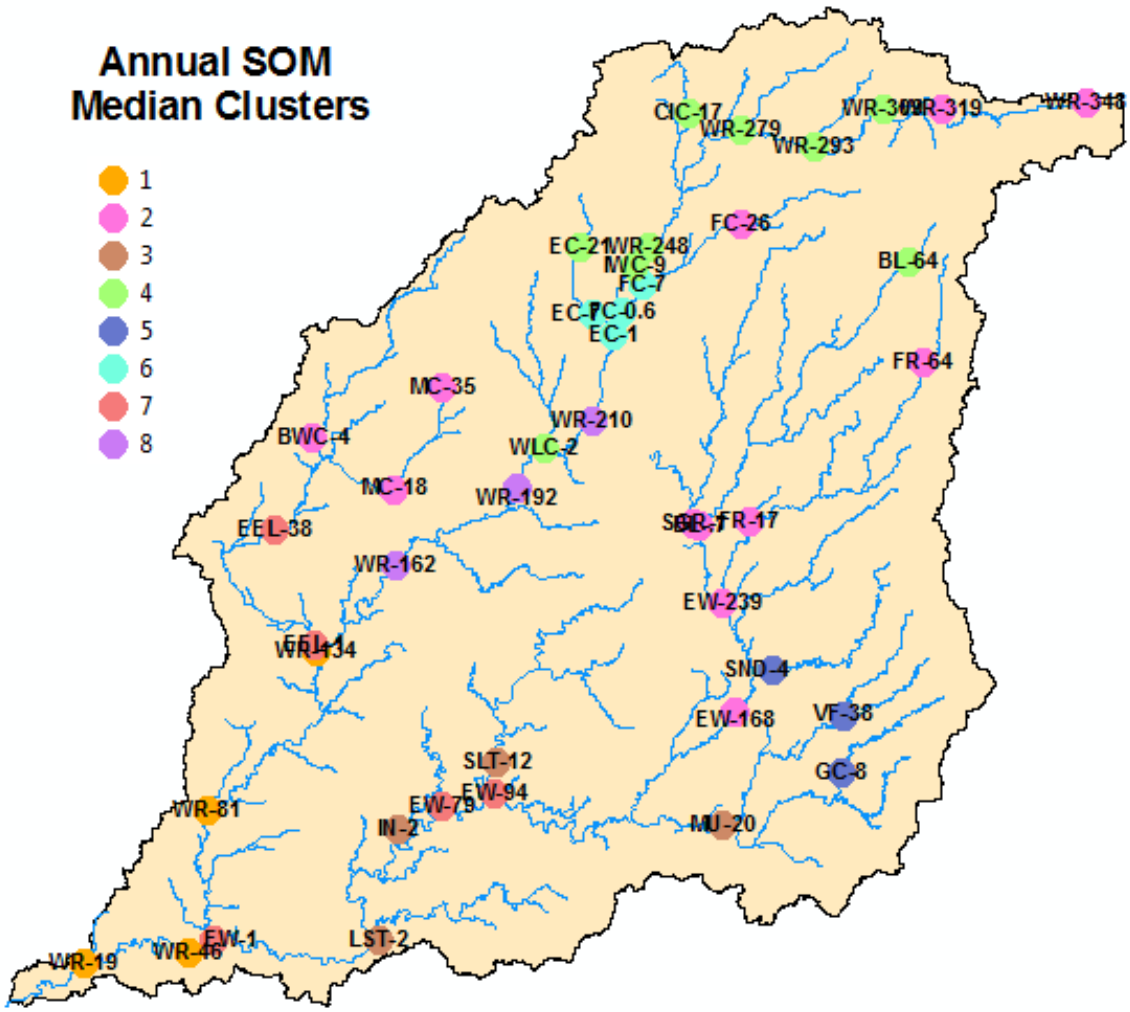
Supplementary Figure 6.3 – Spatially distributed clustering for the Annual Trimmed Mean factors (EEL-1 belongs to cluster 3 and WR-134 belongs to cluster 5; SGR-1 belongs to cluster 5 and BL-.7 belongs to cluster 1)



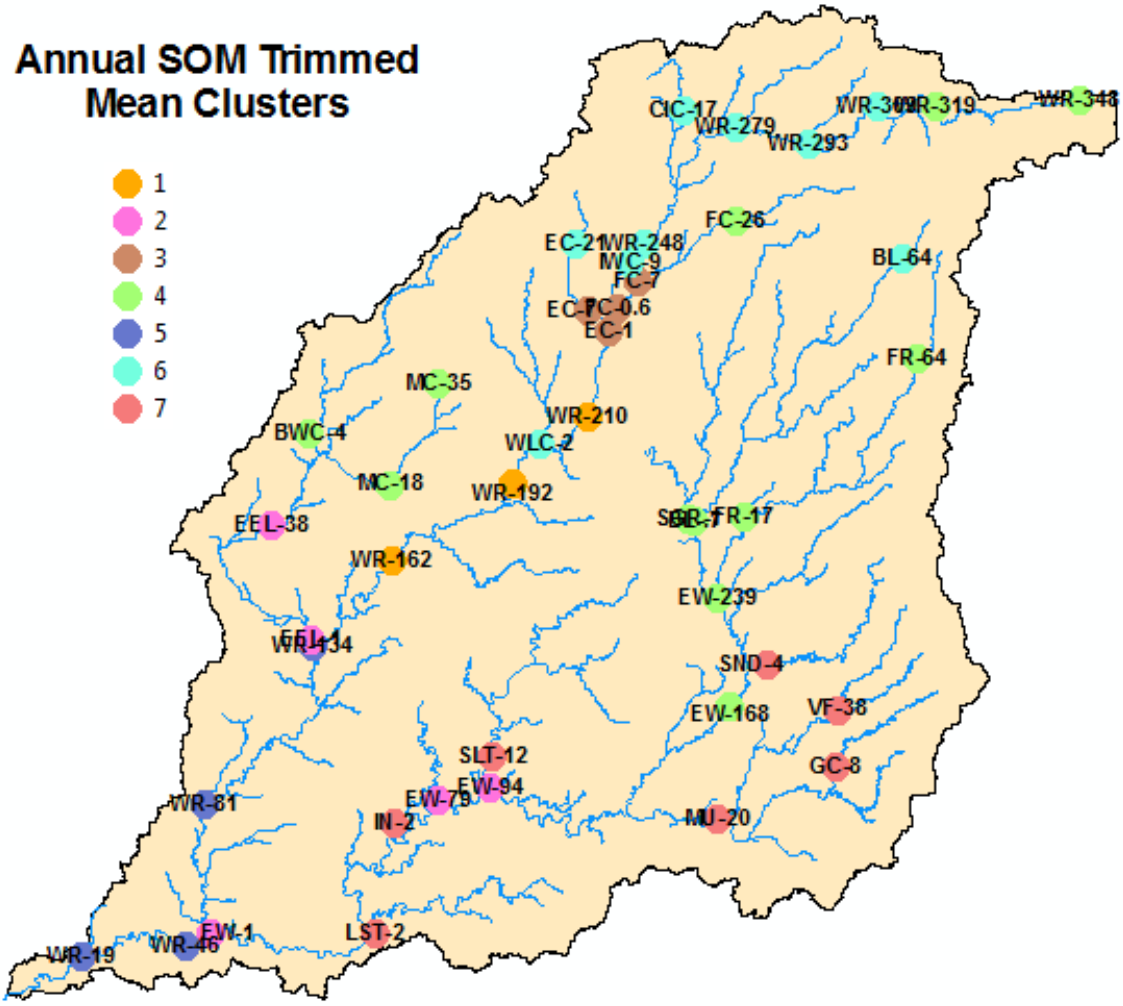
Supplementary Figure 6.4 – Spatially distributed clustering for the Annual Geometric Mean factors (EEL-1 belongs to cluster 3 and WR-134 belongs to cluster 1; SGR-1 and BL-.7 belong to cluster 5)



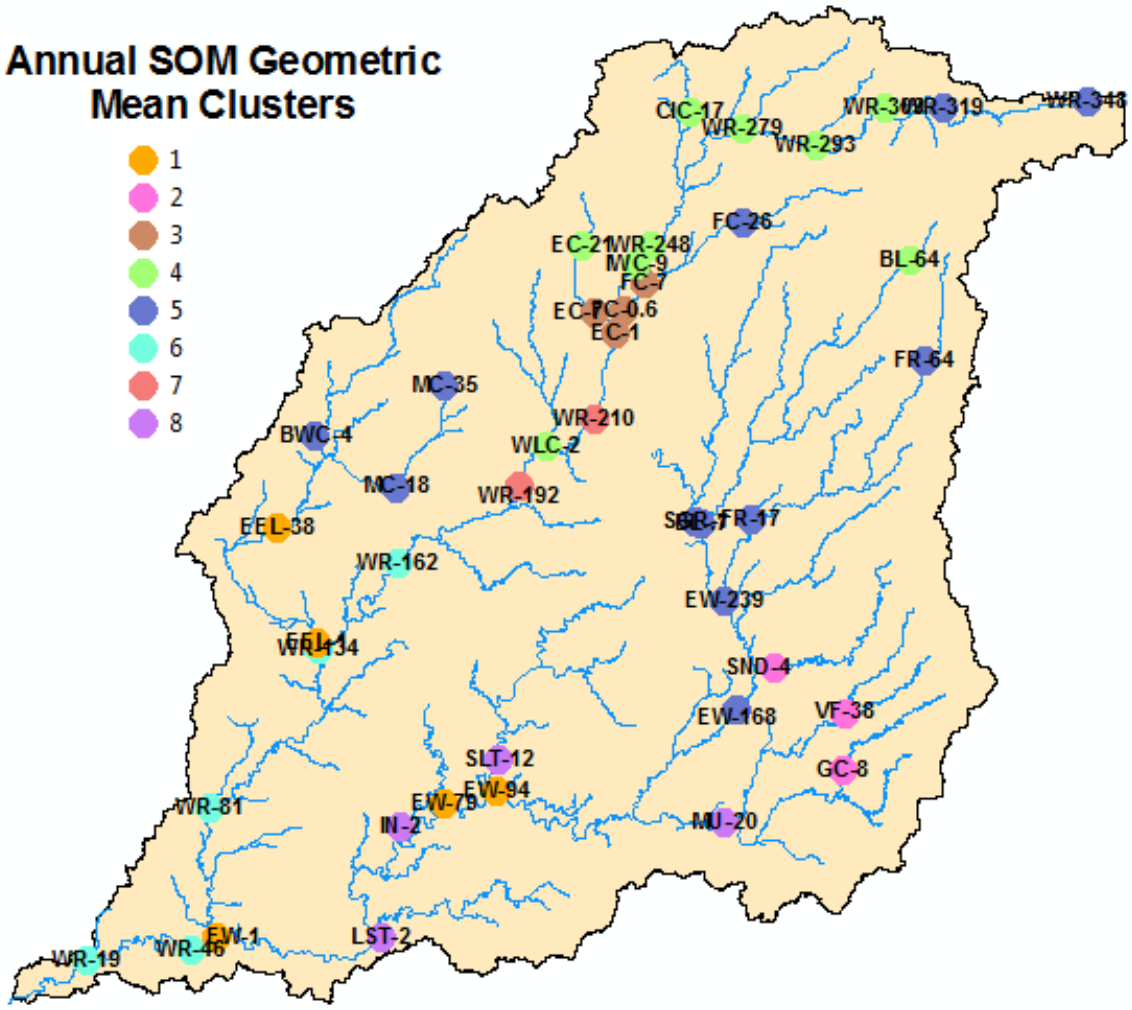
Supplementary Figure 6.5 – Spatially distributed clustering for the Annual Mean SOM (EEL-1 belongs to cluster 2 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belong to cluster 1)



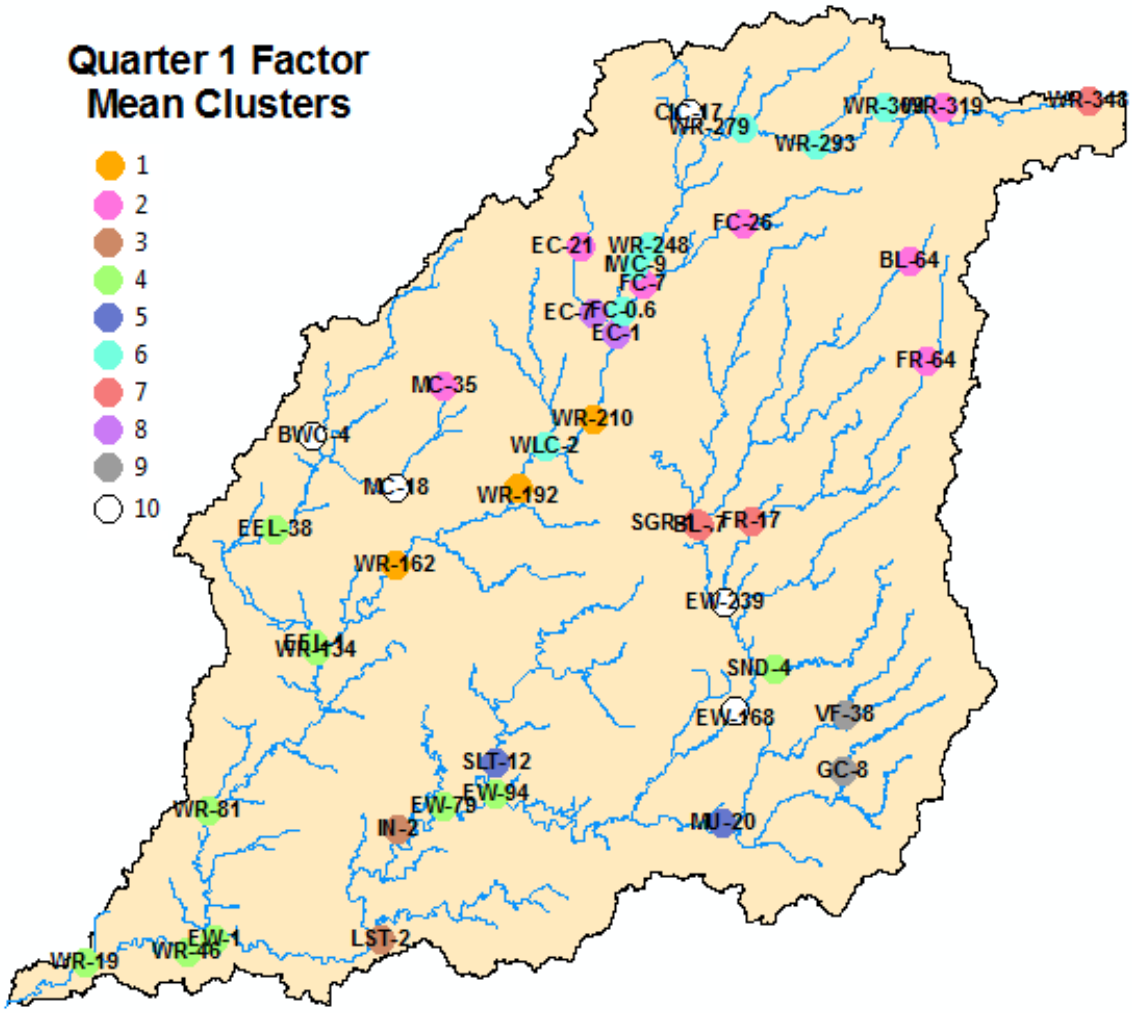
Supplementary Figure 6.6 – Spatially distributed clustering for the Annual Median SOM (EEL-1 belongs to cluster 4 and WR-134 belongs to cluster 7; SGR-1 and BL-.7 belong to cluster 1)



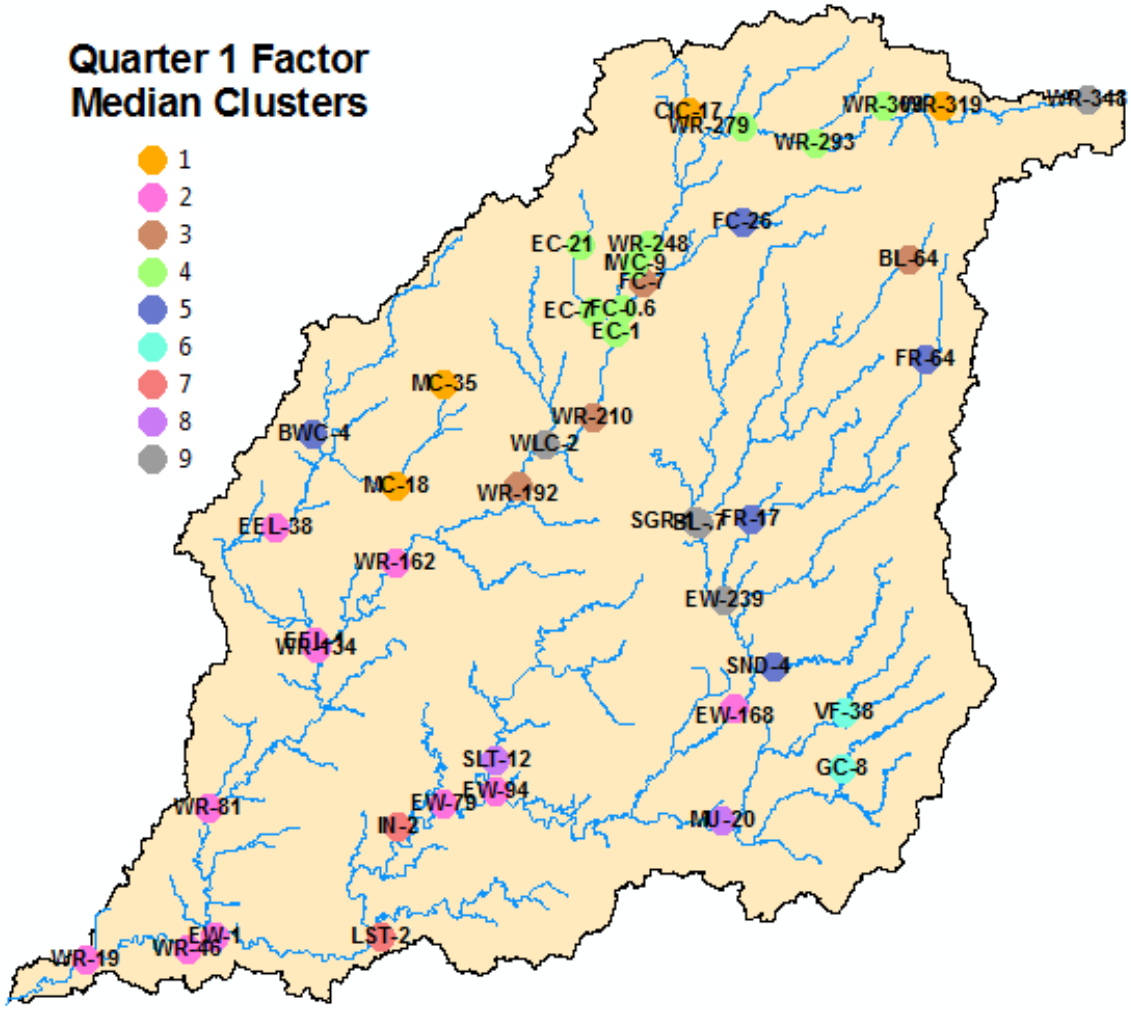
Supplementary Figure 6.7 – Spatially distributed clustering for the Annual Trimmed Mean SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belong to cluster 2)



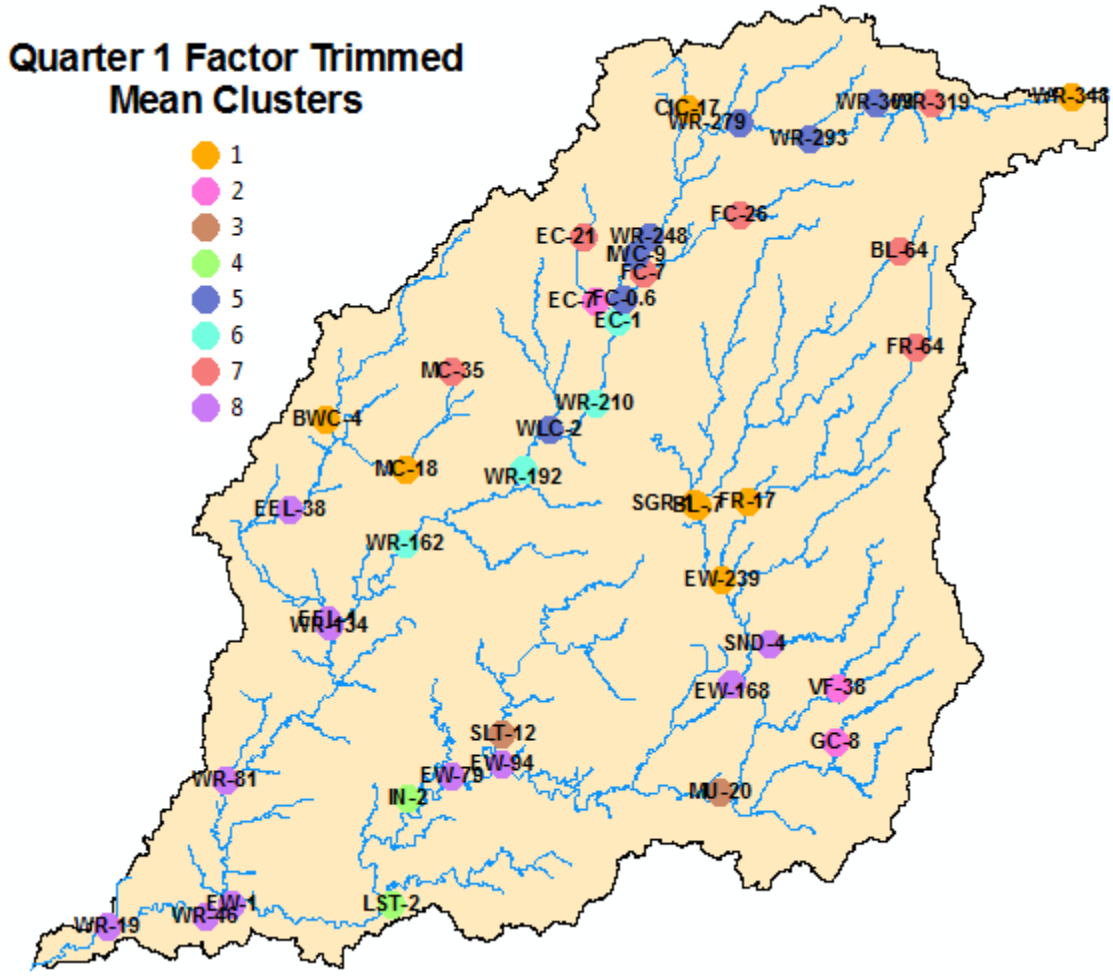
Supplementary Figure 6.8 – Spatially distributed clustering for the Annual Geometric Mean SOM (EEL-1 belongs to cluster 1 and WR-134 belongs to cluster 4; SGR-1 and BL-.7 belong to cluster 2)



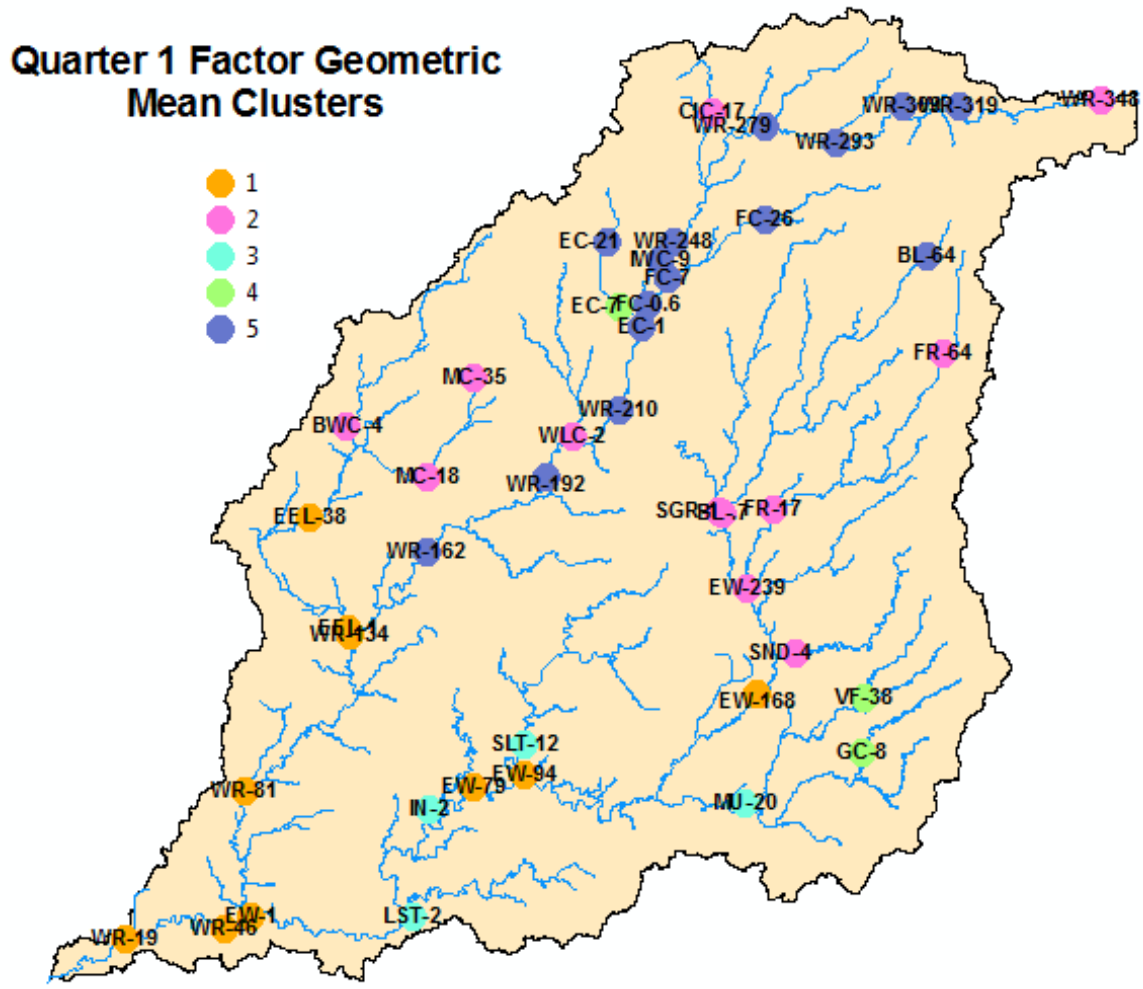
Supplementary Figure 6.9 – Spatially distributed clustering for the Quarter 1 Mean factors (EEL-1 and WR-134 belong to cluster 4; SGR-1 and BL-.7 belong to cluster 7)



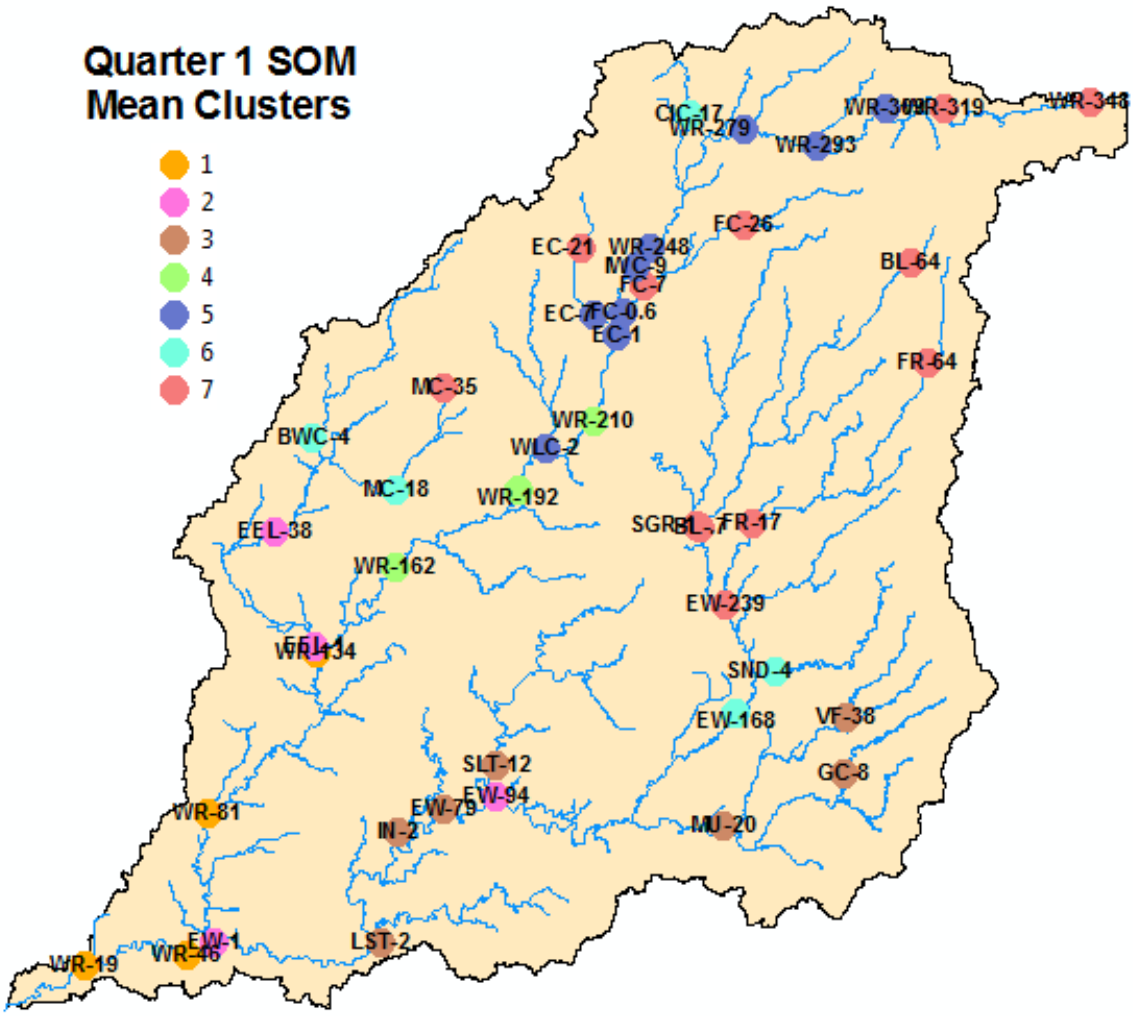
Supplementary Figure 6.10 – Spatially distributed clustering for the Quarter 1 Median factors (EEL-1 and WR-134 belong to cluster 2; SGR-1 and BL-.7 belong to cluster 9)



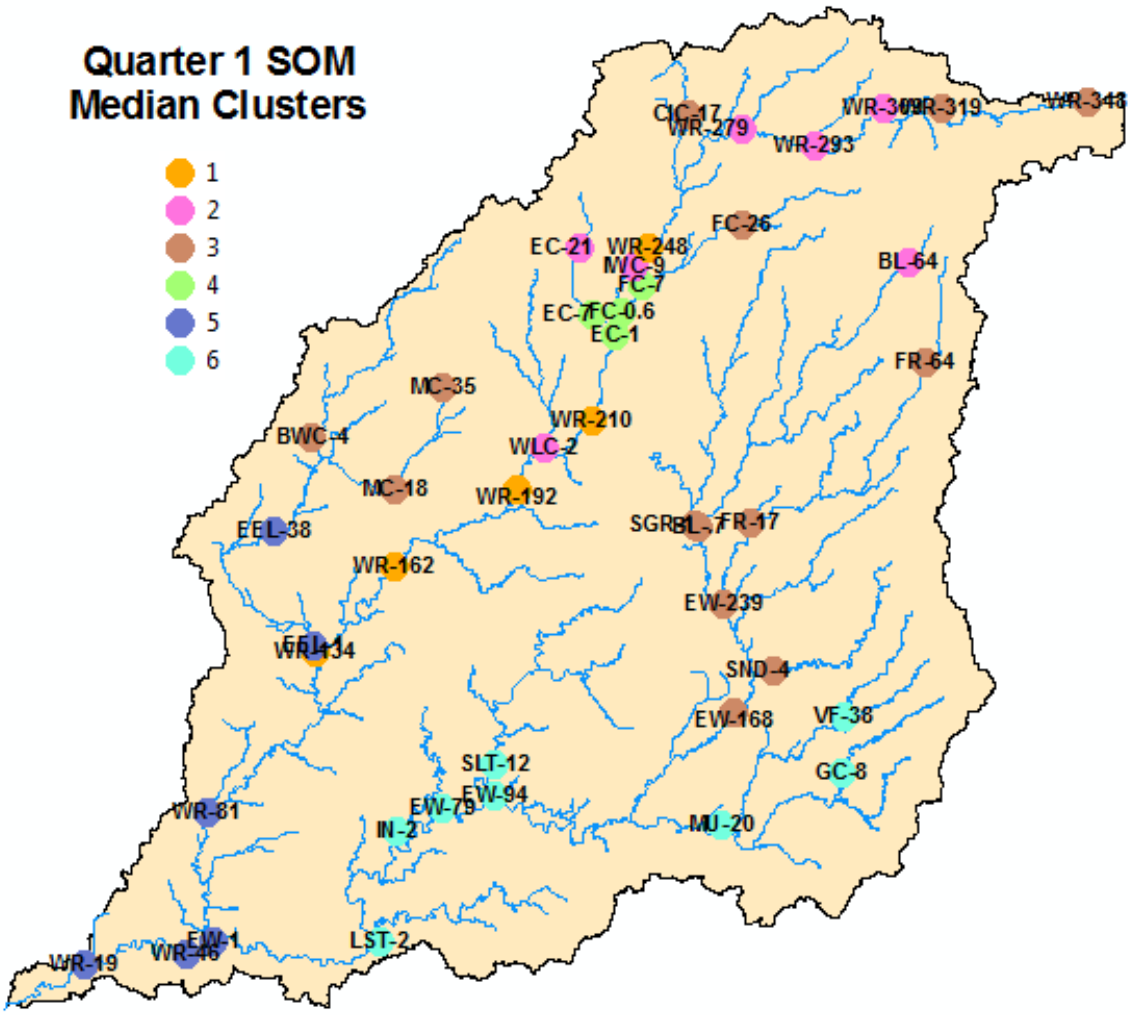
Supplementary Figure 6.11 – Spatially distributed clustering for the Quarter 1 Trimmed Mean factors (EEL-1 and WR-134 belong to cluster 8; SGR-1 and BL-.7 belong to cluster 1)



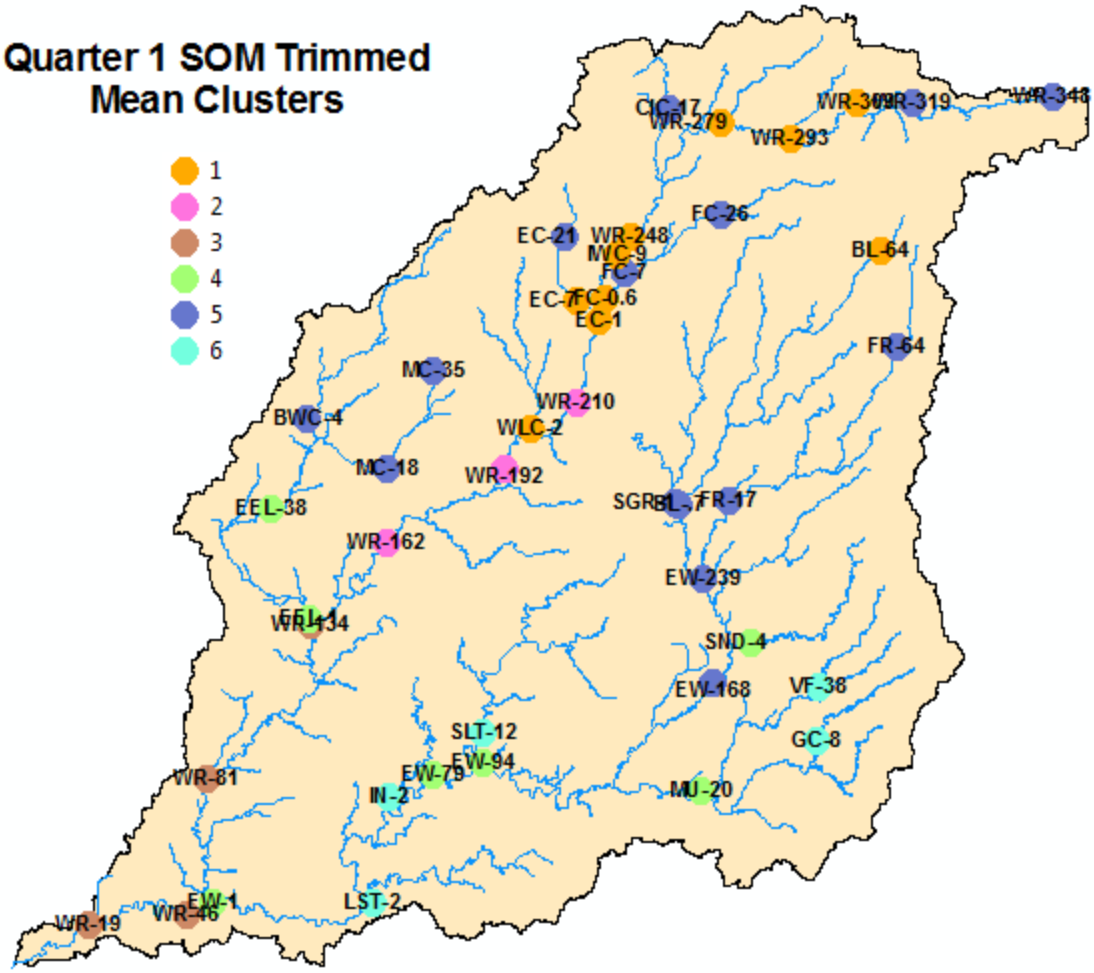
Supplementary Figure 6.12 – Spatially distributed clustering for the Quarter 1 Geometric Mean factors (EEL-1 and WR-134 belong to cluster 1; SGR-1 and BL-.7 belong to cluster 2)



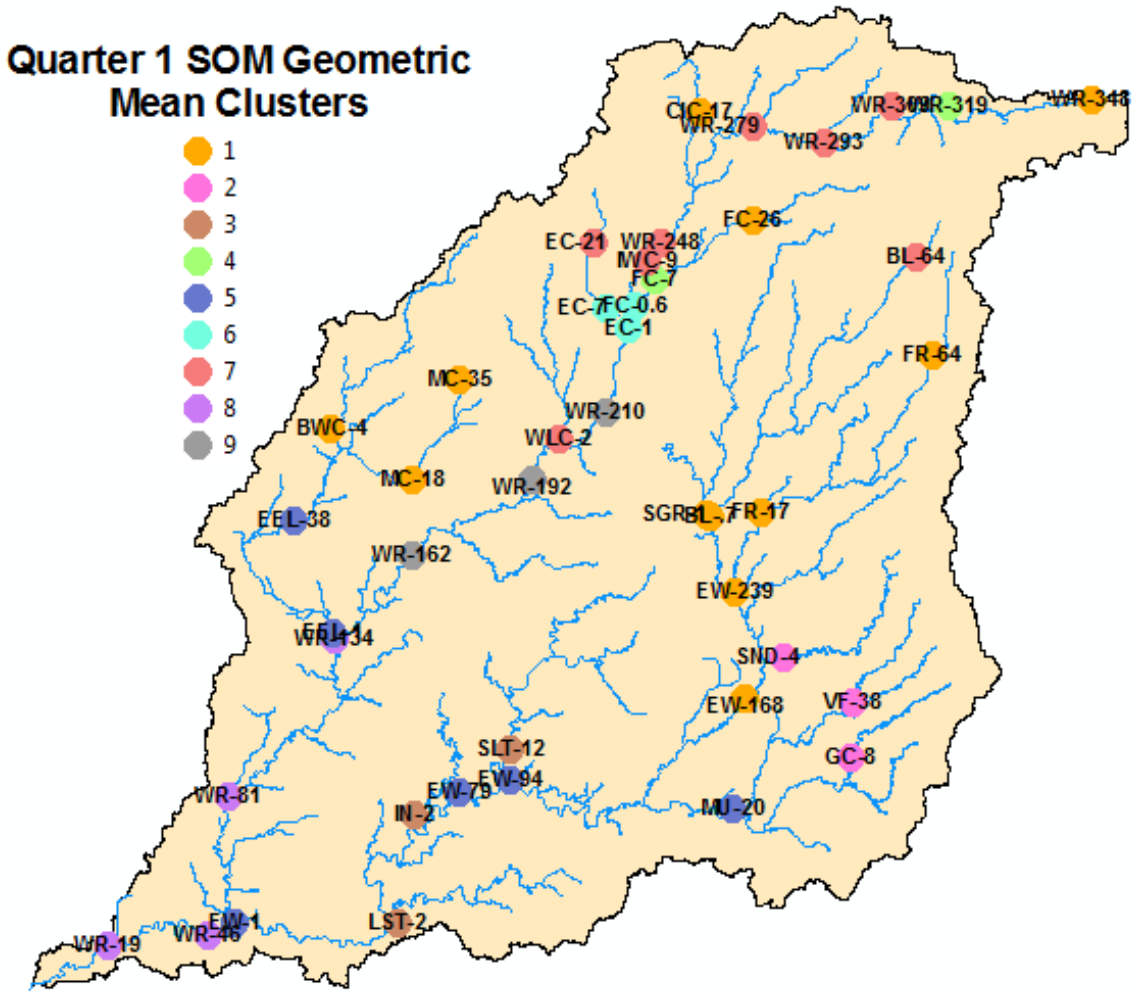
Supplementary Figure 6.13 – Spatially distributed clustering for the Quarter 1 Mean SOM (EEL-1 and belongs to cluster 2 WR-134 belong to cluster 1; SGR-1 and BL-.7 belong to cluster 7)



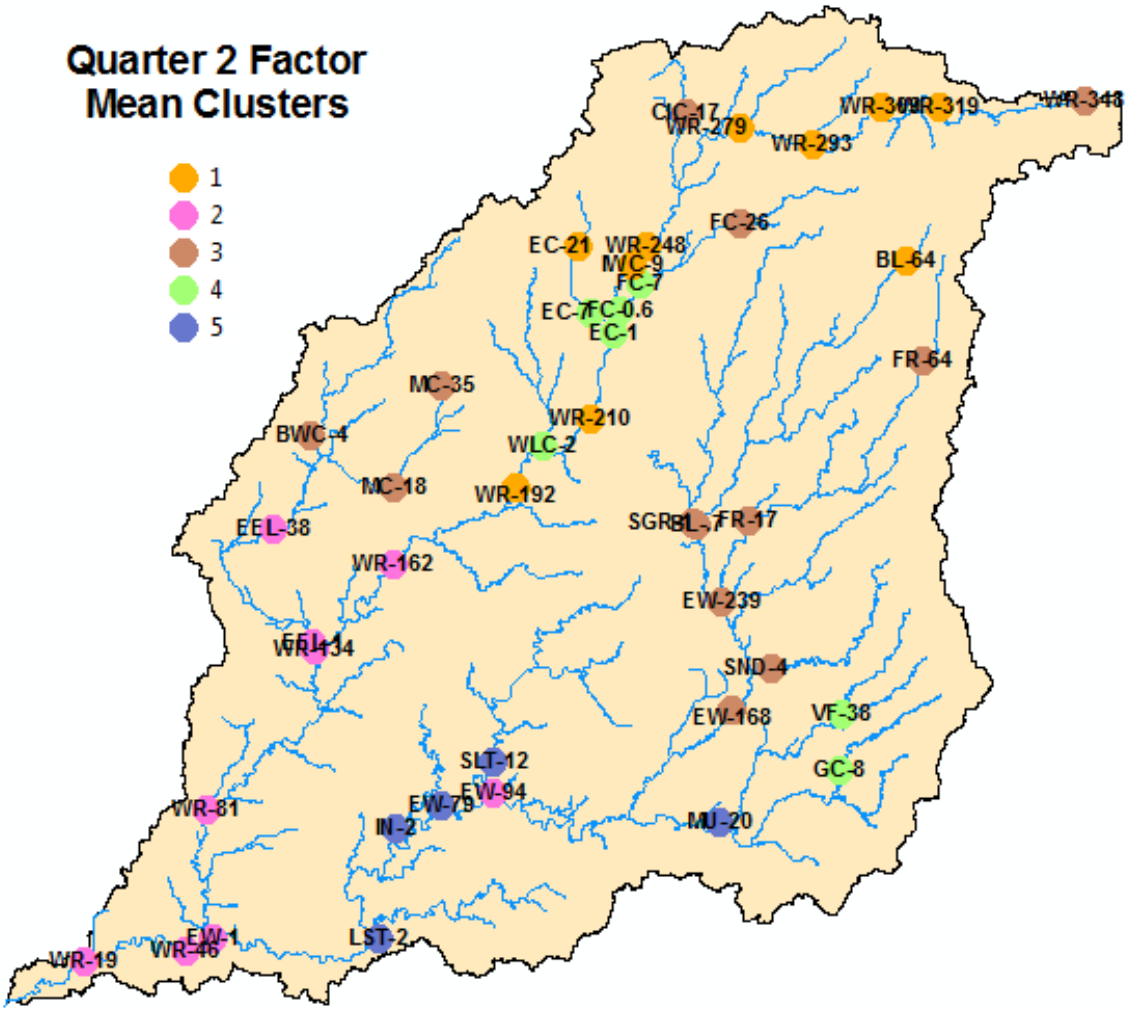
Supplementary Figure 6.14 – Spatially distributed clustering for the Quarter 1 Median SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 1; SGR-1 and BL-.7 belong to cluster 3)



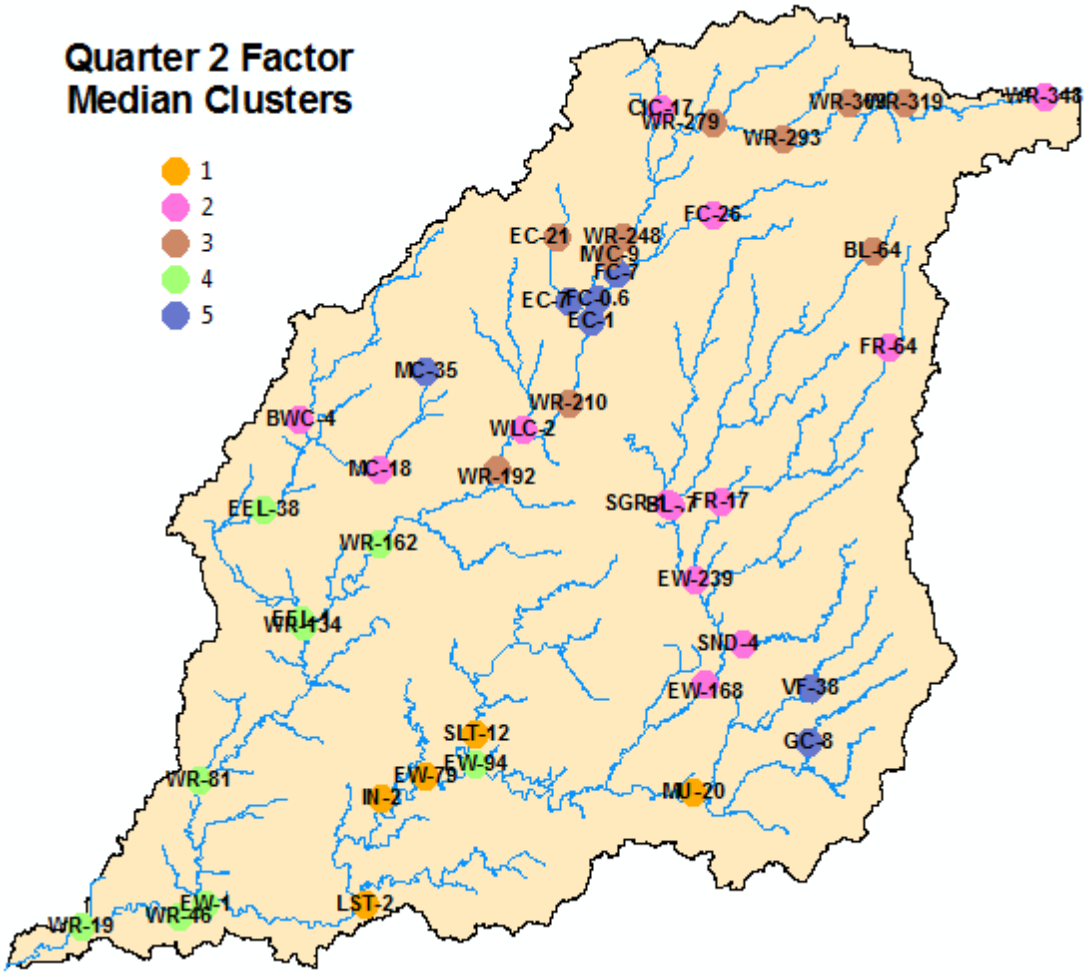
Supplementary Figure 6.15 – Spatially distributed clustering for the Quarter 1 TrImmed Mean SOM (EEL-1 belongs to cluster 4 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belong to cluster 5)



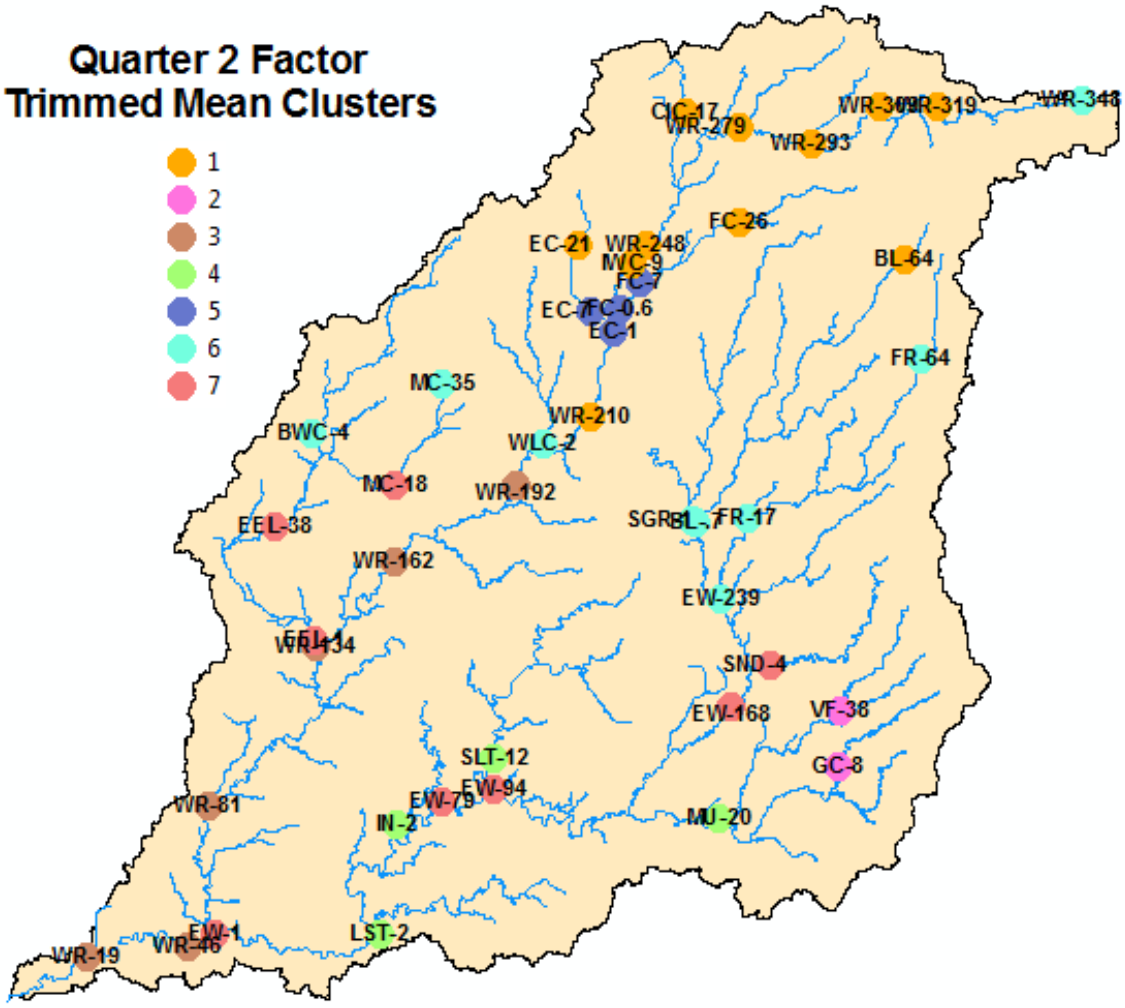
Supplementary Figure 6.16 – Spatially distributed clustering for the Quarter 1 Geometric Mean SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 8; SGR-1 and BL-.7 belong to cluster 1)



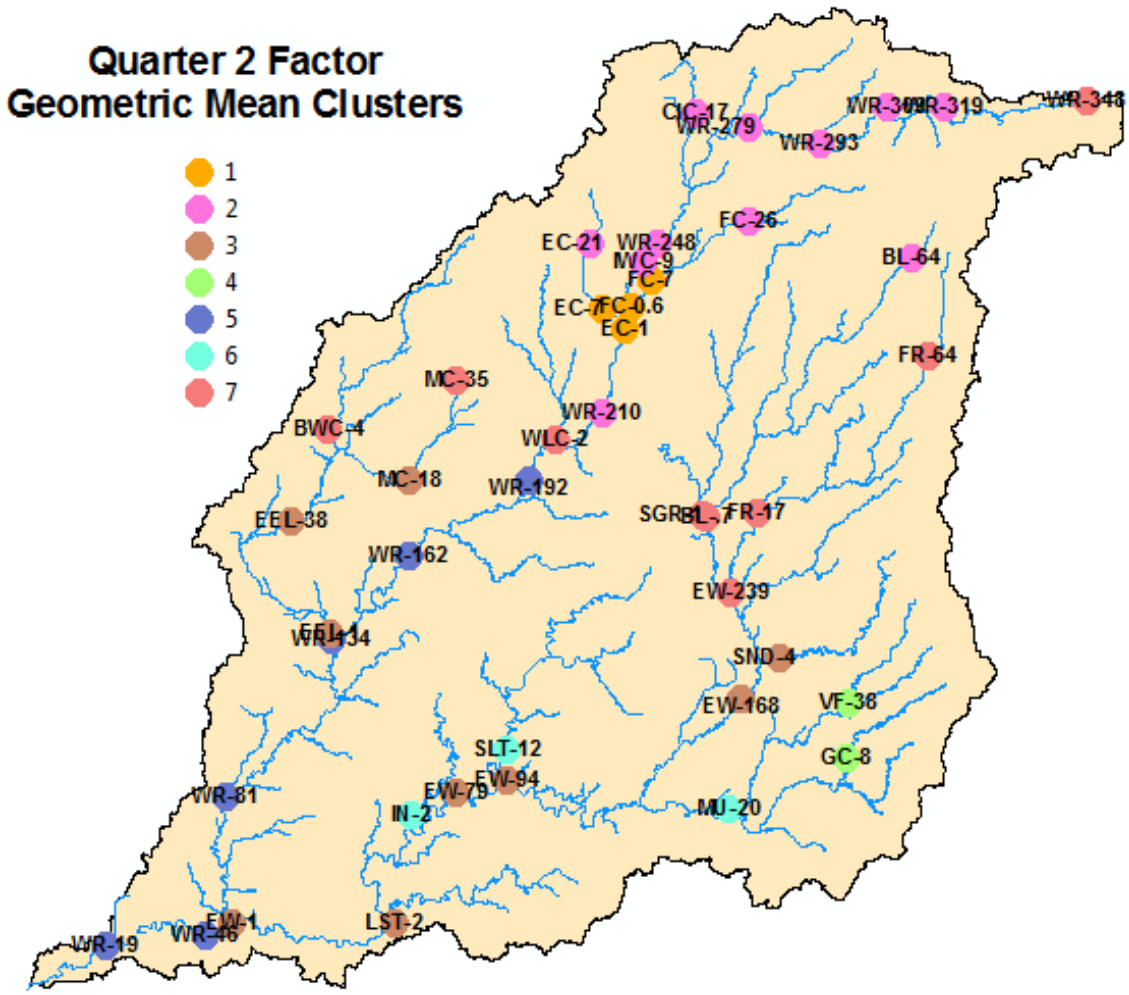
Supplementary Figure 6.17 – Spatially distributed clustering for the Quarter 2 Mean factors (EEL-1 and WR-134 belong to cluster 2; SGR-1 and BL-.7 belong to cluster 3)



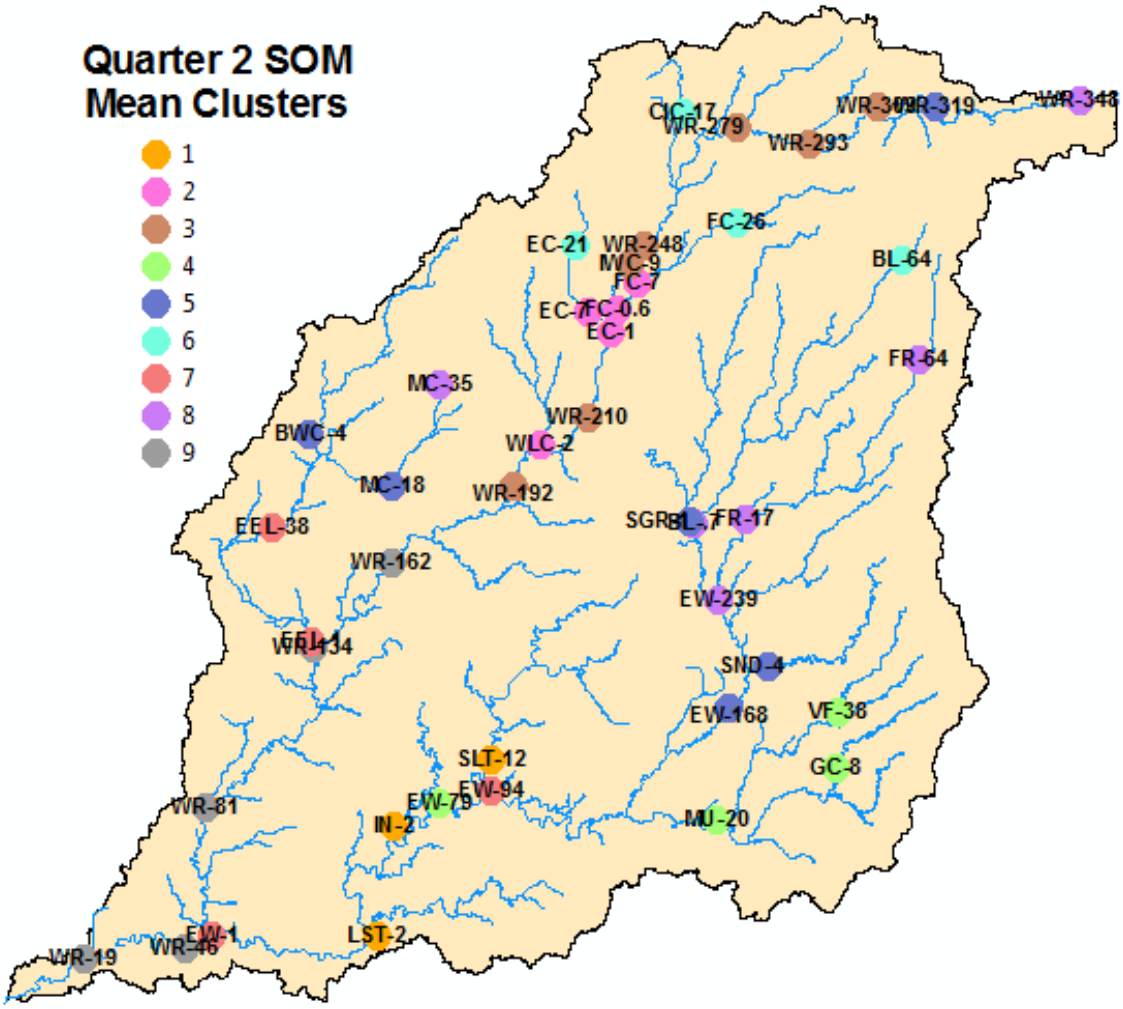
Supplementary Figure 6.18 – Spatially distributed clustering for the Quarter 2 Median factors (EEL-1 and WR-134 belong to cluster 4; SGR-1 and BL-.7 belong to cluster 2)



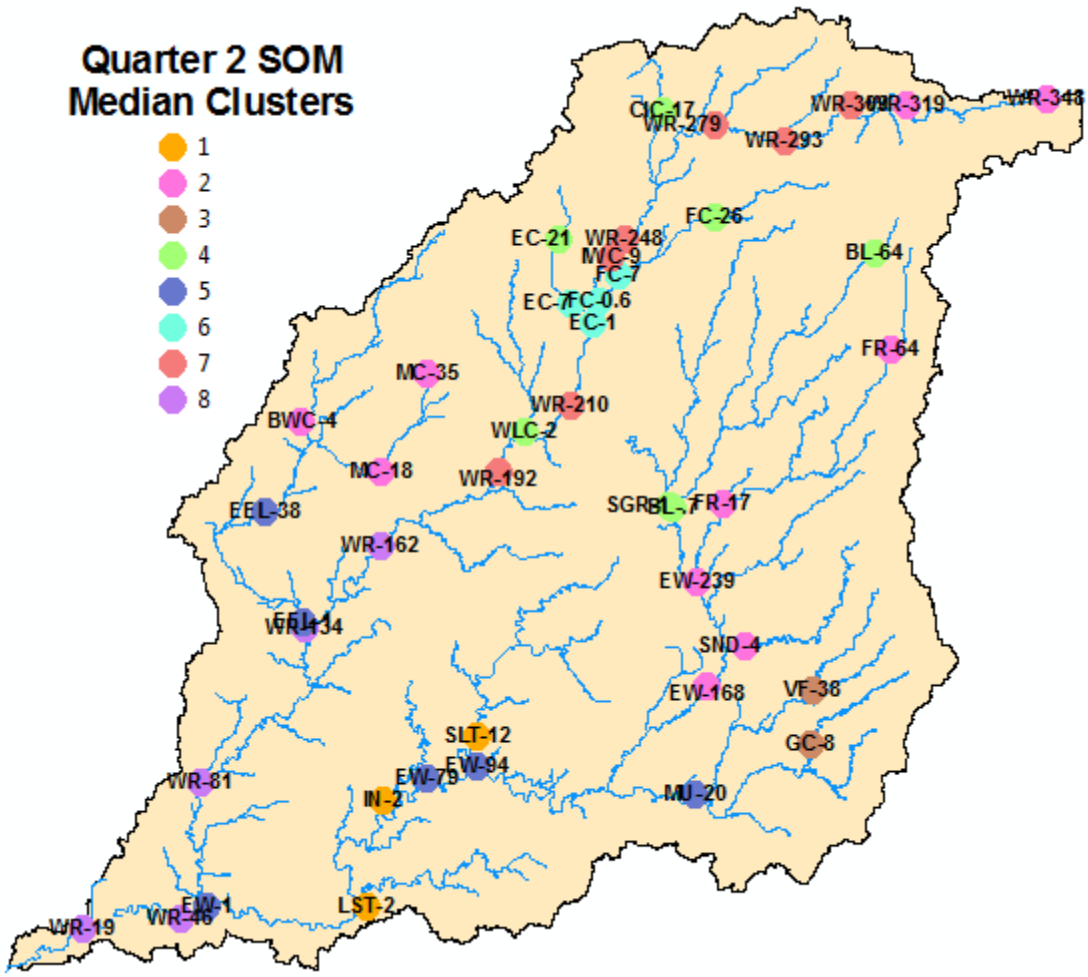
Supplementary Figure 6.19 – Spatially distributed clustering for the Quarter 2 Trimmed Mean factors (EEL-1 belongs to cluster 7 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belong to cluster 6)



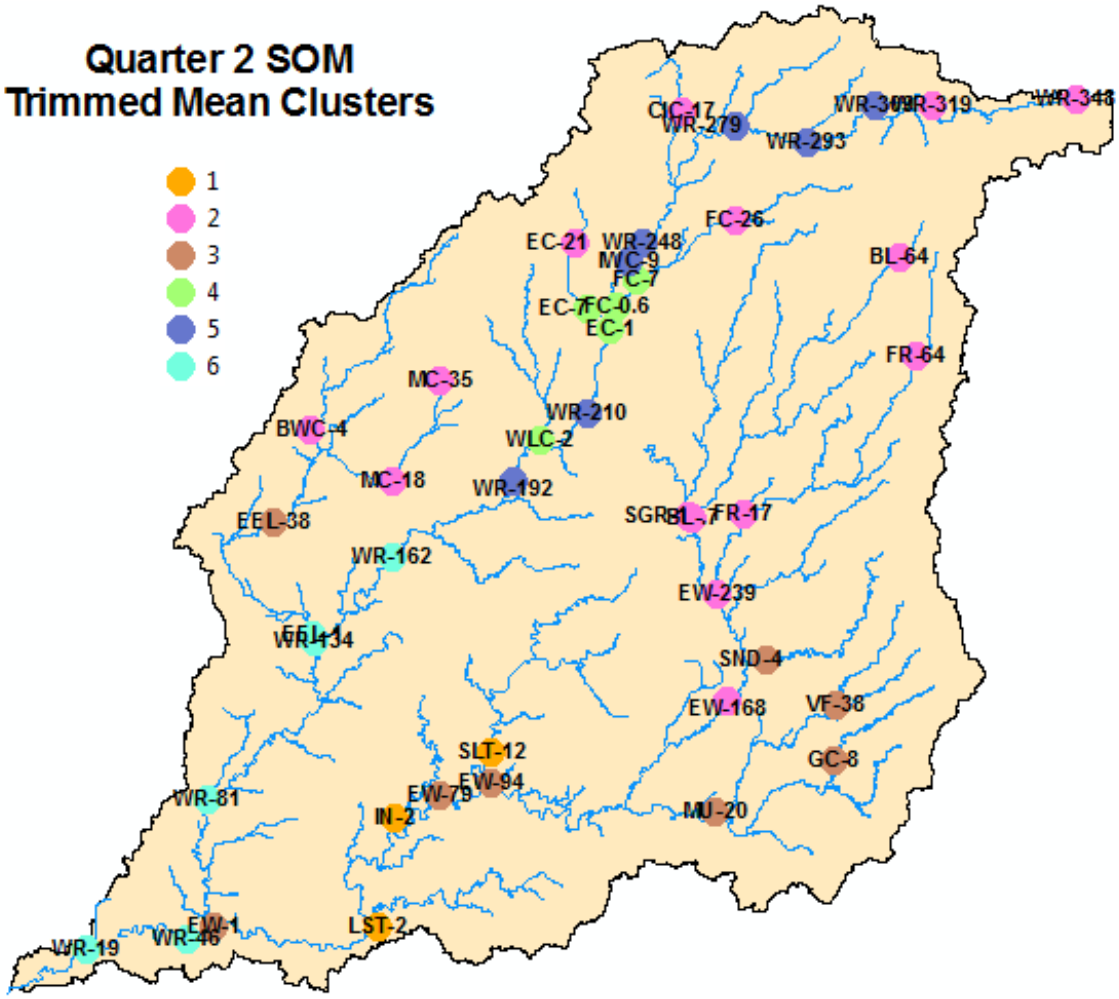
Supplementary Figure 6. 20– Spatially distributed clustering for the Quarter 2 Geometric Mean factors (EEL-1 belongs to cluster 3 and WR-134 belongs to cluster 5; SGR-1 and BL-.7 belong to cluster 7)



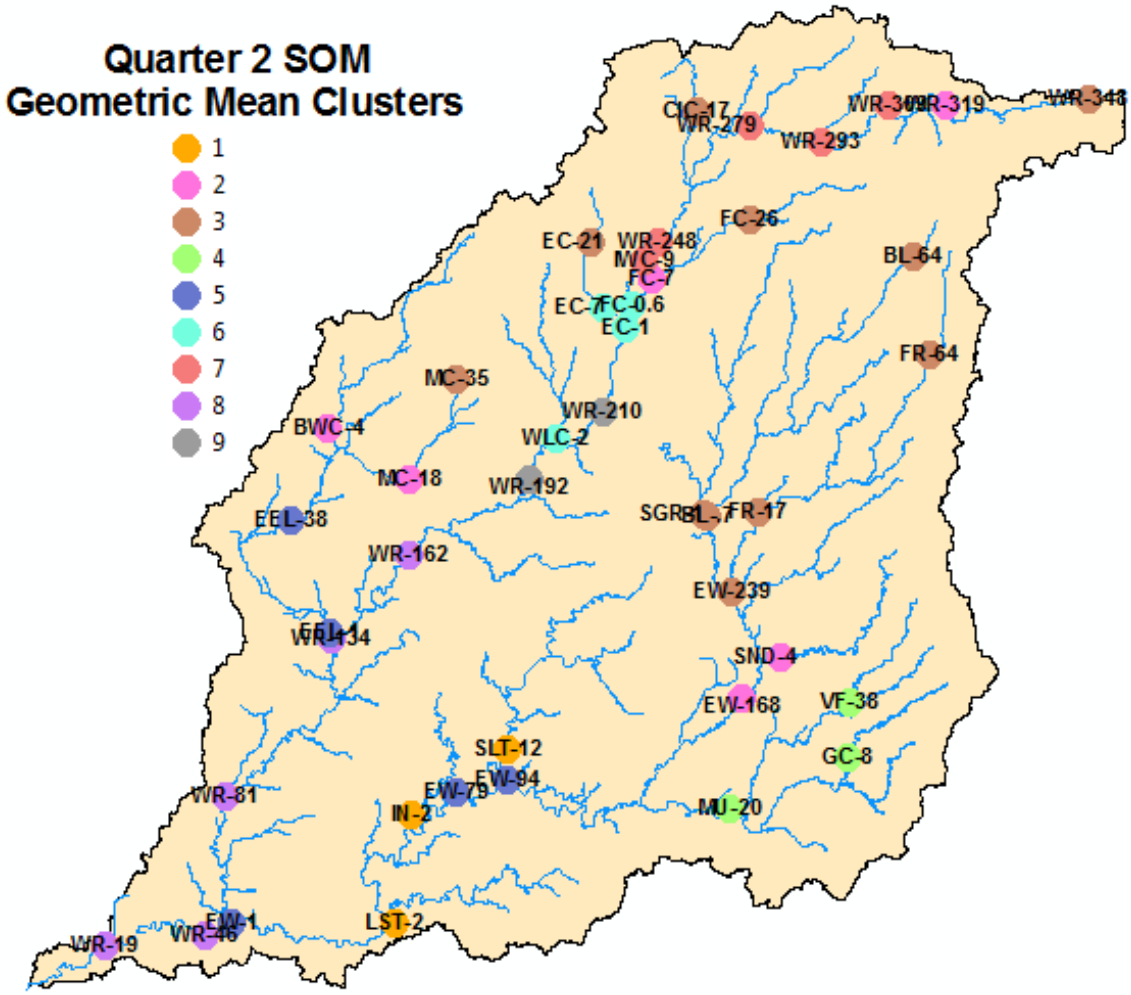
Supplementary Figure 6.21 – Spatially distributed clustering for the Quarter 2 Mean SOM (EEL-1 belongs to cluster 7 and WR-134 belongs to cluster 9; SGR-1 belongs to cluster 5 and BL-.7 belongs to cluster 8)



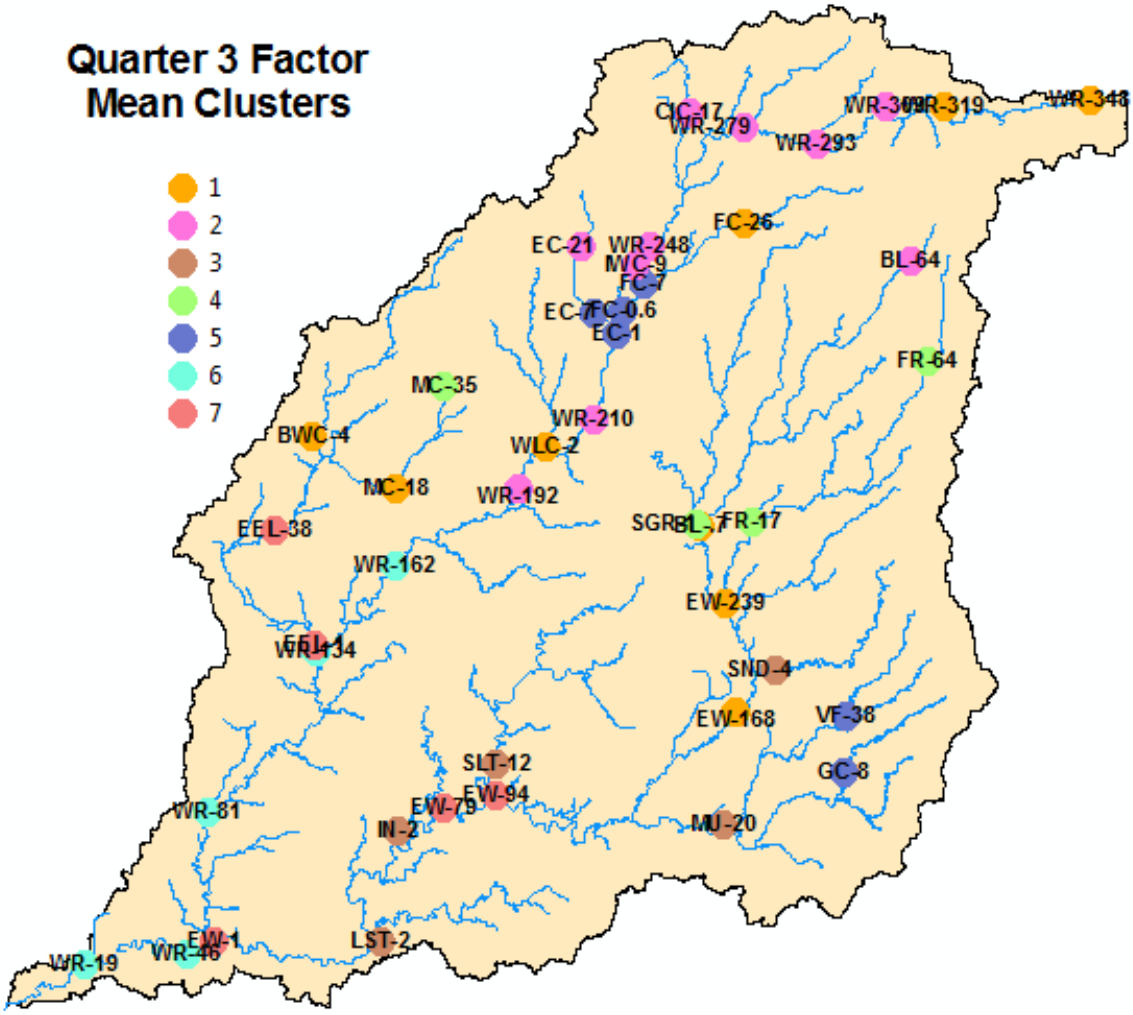
Supplementary Figure 6.22 – Spatially distributed clustering for the Quarter 2 Median SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 8; SGR-1 and BL-.7 belong to cluster 4)



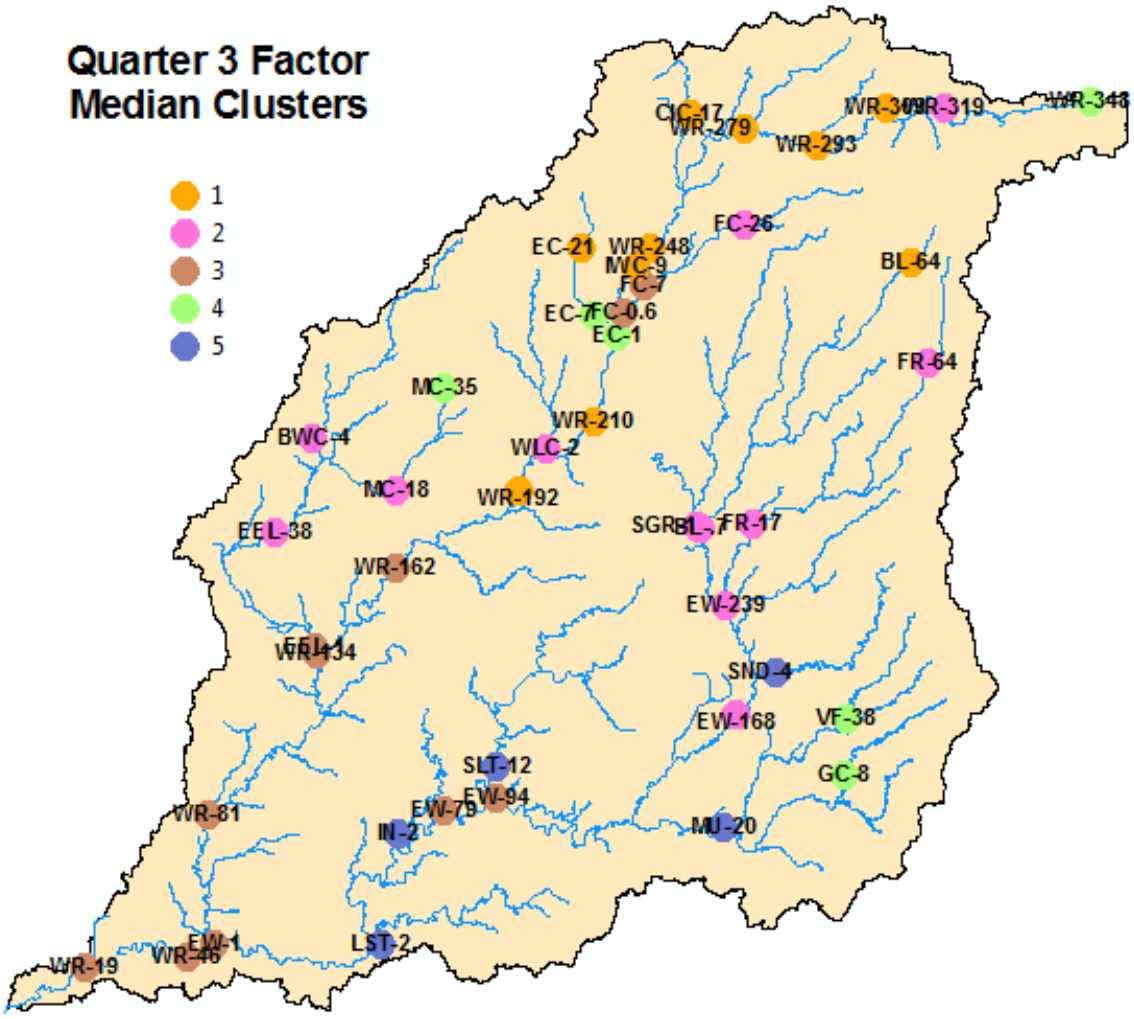
Supplementary Figure 6.23 – Spatially distributed clustering for the Quarter 2 Trimmed Mean SOM (EEL-1 and WR-134 belong to cluster 6; SGR-1 and BL-.7 belong to cluster 2)



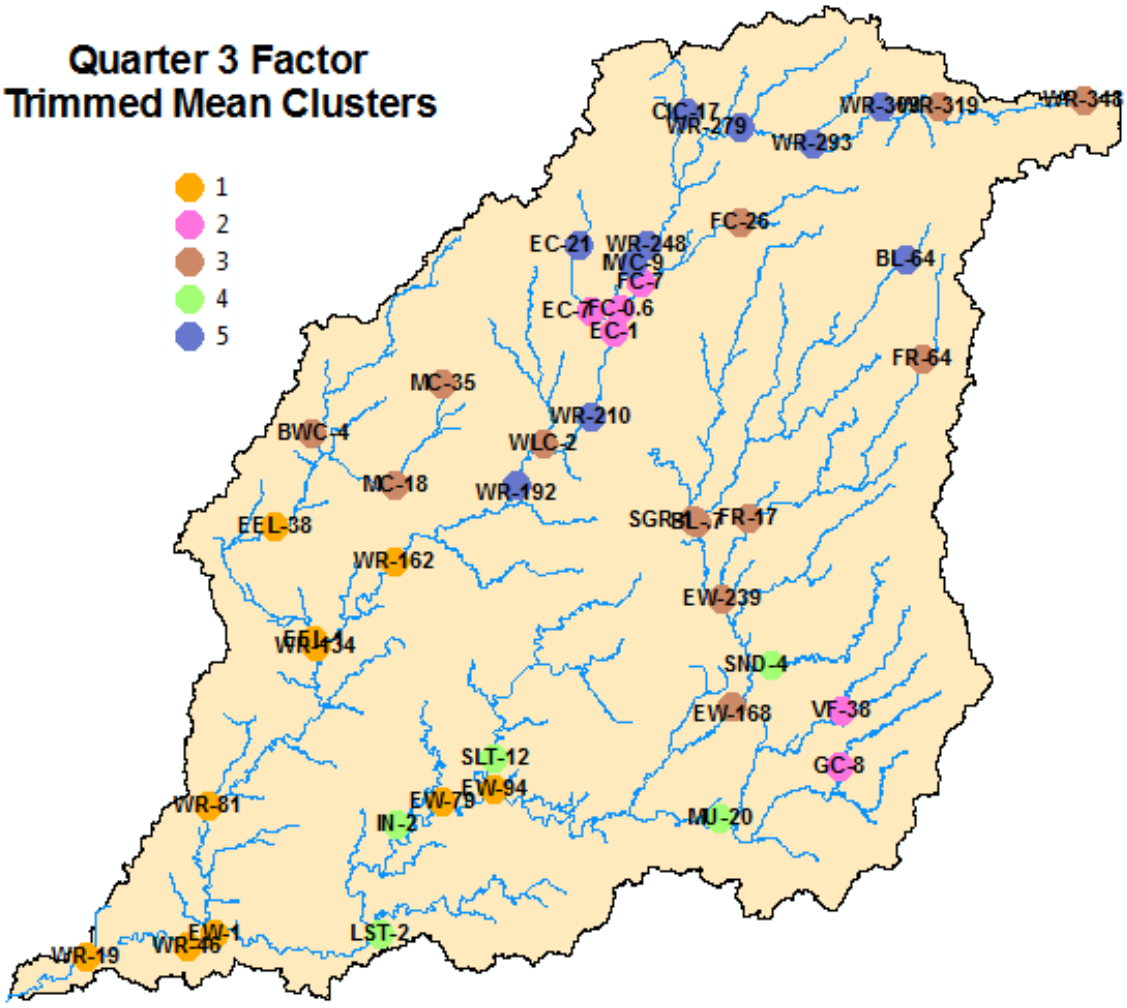
Supplementary Figure 6.24 – Spatially distributed clustering for the Quarter 2 Geometric Mean SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 8; SGR-1 and BL-.7 belong to cluster 3)



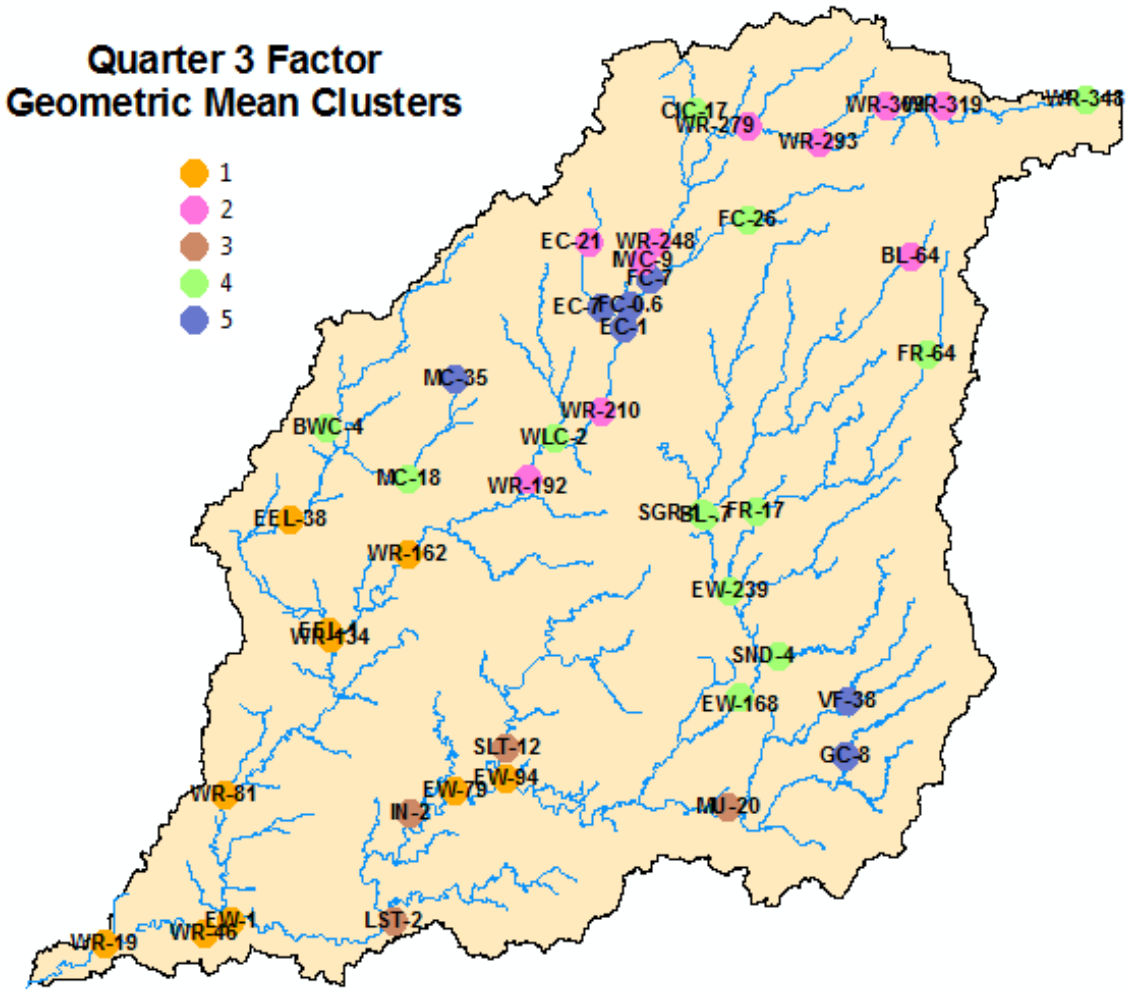
Supplementary Figure 6.25 – Spatially distributed clustering for the Quarter 3 Mean factors (EEL-1 belongs to cluster 7 and WR-134 belongs to cluster 6; SGR-1 belongs to cluster 4 and BL-.7 belongs to cluster 1)



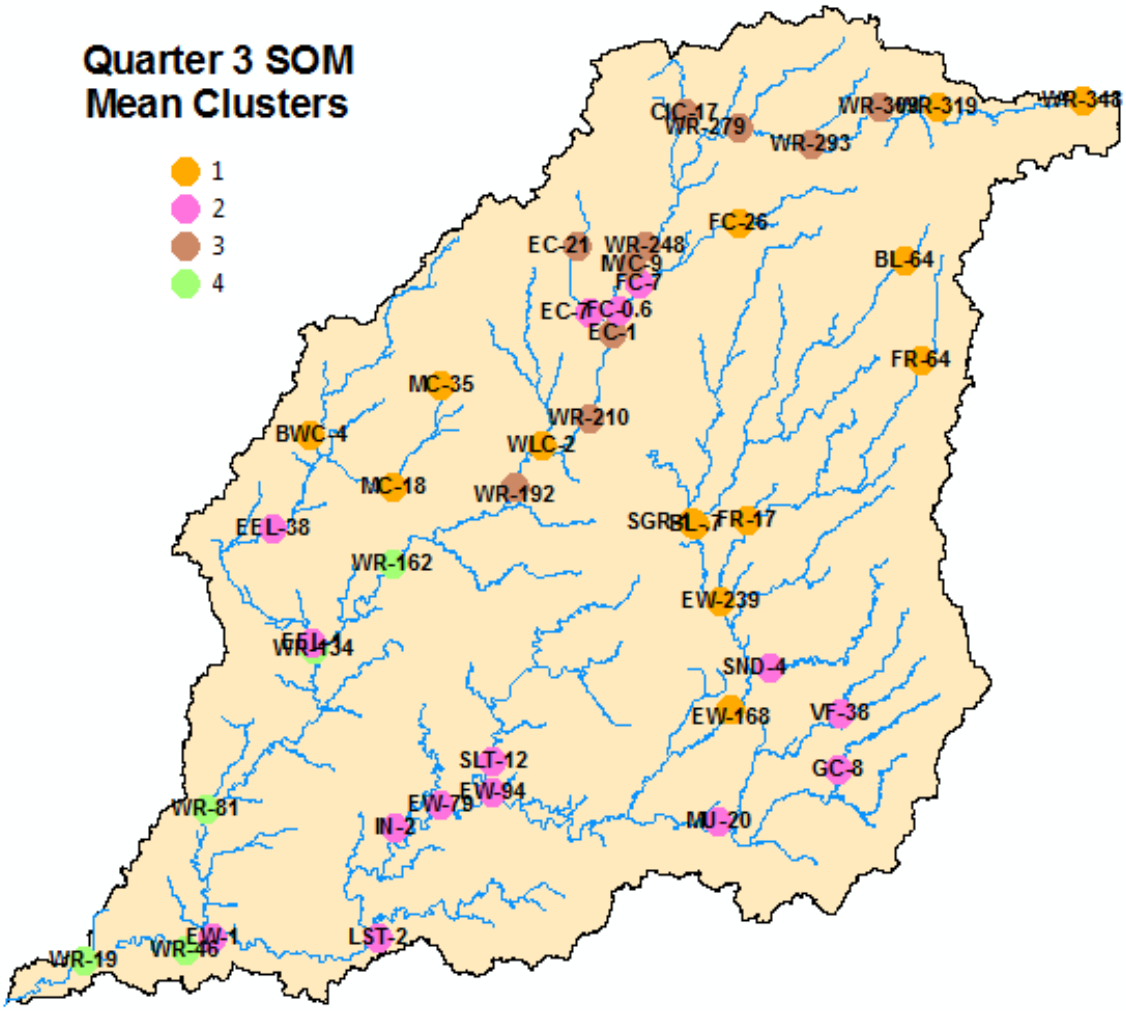
Supplementary Figure 6.26 – Spatially distributed clustering for the Quarter 3 Median factors (EEL-1 and WR-134 belong to cluster 3; SGR-1 and BL-.7 belong to cluster 2)



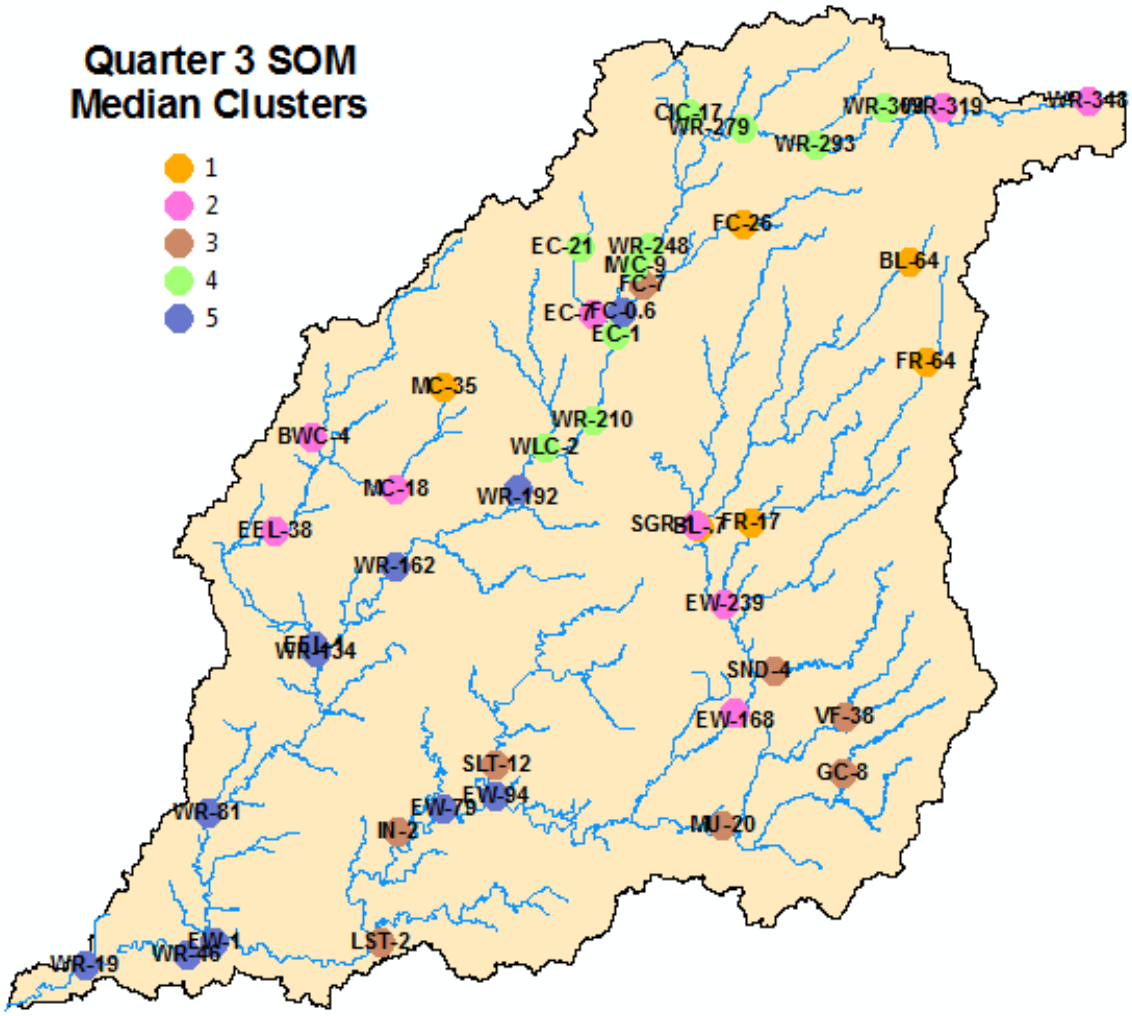
Supplementary Figure 6.27 – Spatially distributed clustering for the Quarter 3 Trimmed Mean factors (EEL-1 and WR-134 belong to cluster 1; SGR-1 and BL-.7 belong to cluster 3)



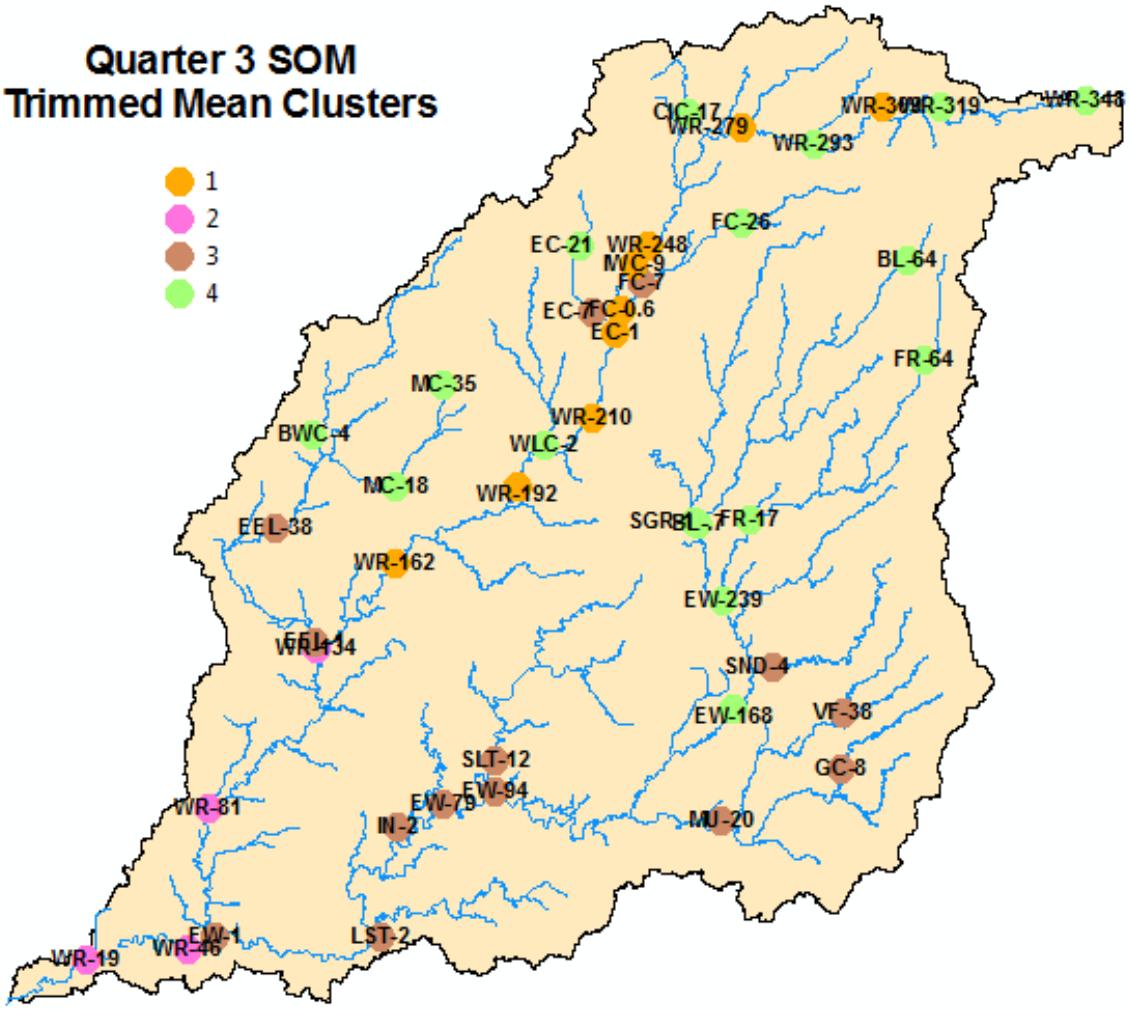
Supplementary Figure 6.28 – Spatially distributed clustering for the Quarter 3 Geometric Mean factors (EEL-1 and WR-134 belong to cluster 1; SGR-1 and BL-.7 belong to cluster 4)



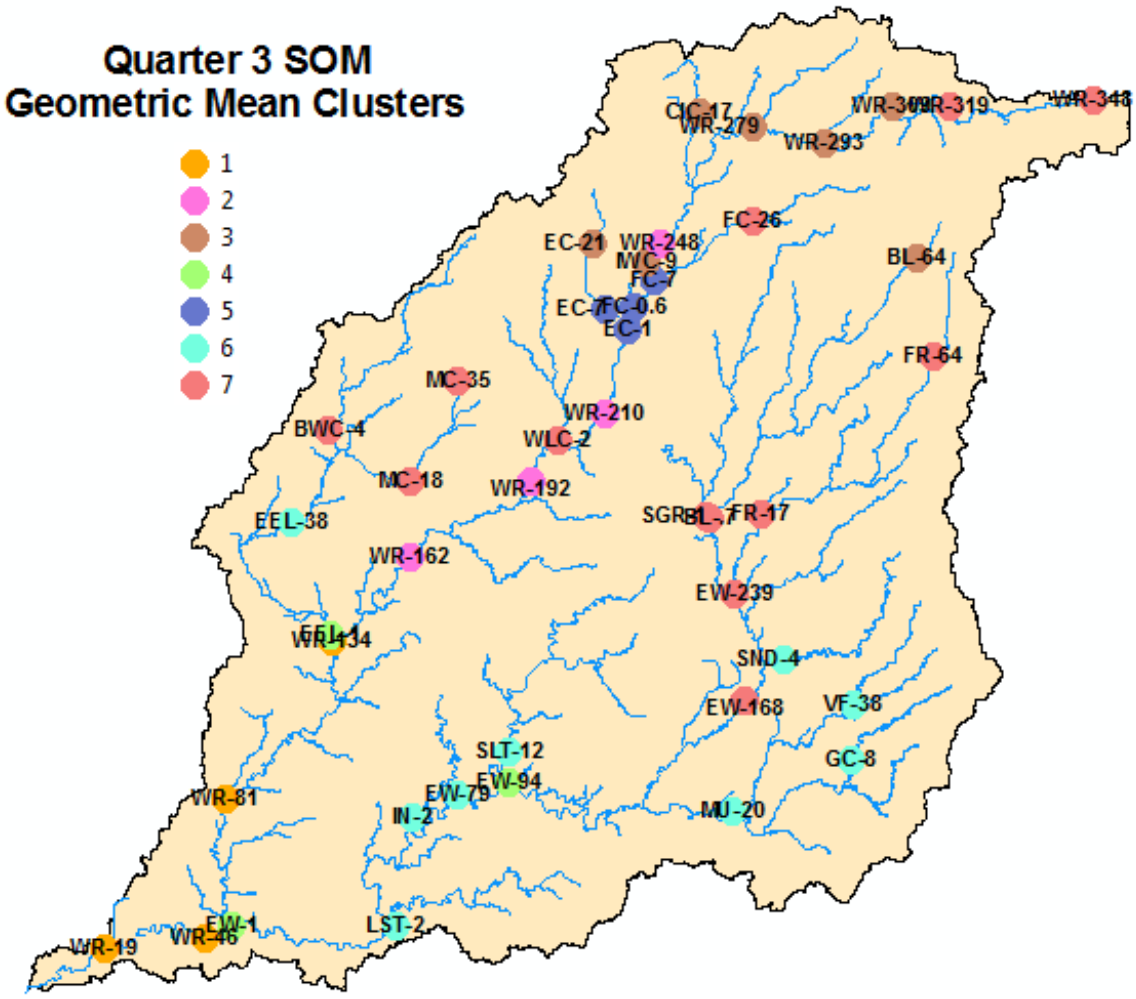
Supplementary Figure 6.29 – Spatially distributed clustering for the Quarter 3 Mean SOM (EEL-1 belongs to cluster 2 and WR-134 belongs to cluster 4; SGR-1 and BL-.7 belong to cluster 1)



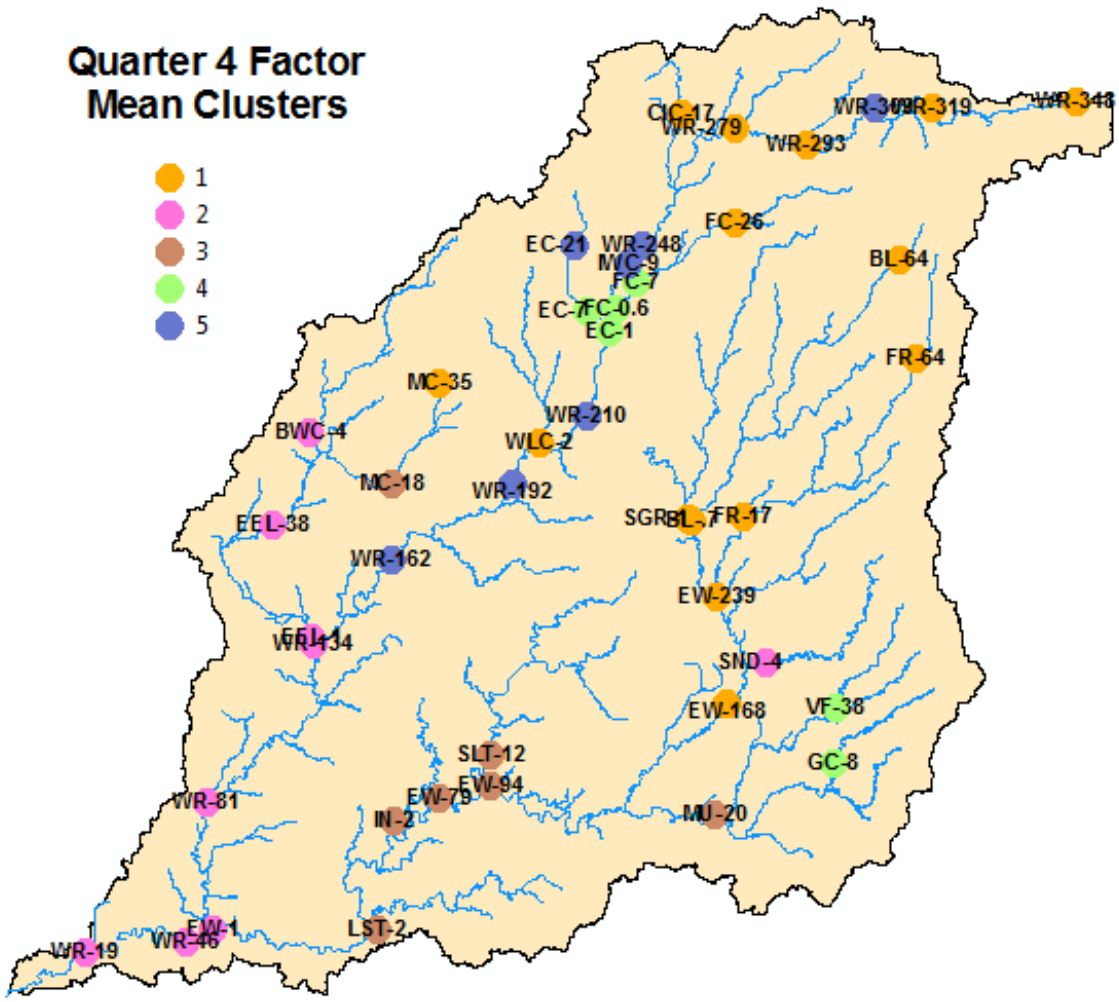
Supplementary Figure 6.30 – Spatially distributed clustering for the Quarter 3 Median SOM (EEL-1 and WR-134 belong to cluster 5; SGR-1 belongs to cluster 2 and BL-.7 belongs to cluster 1)



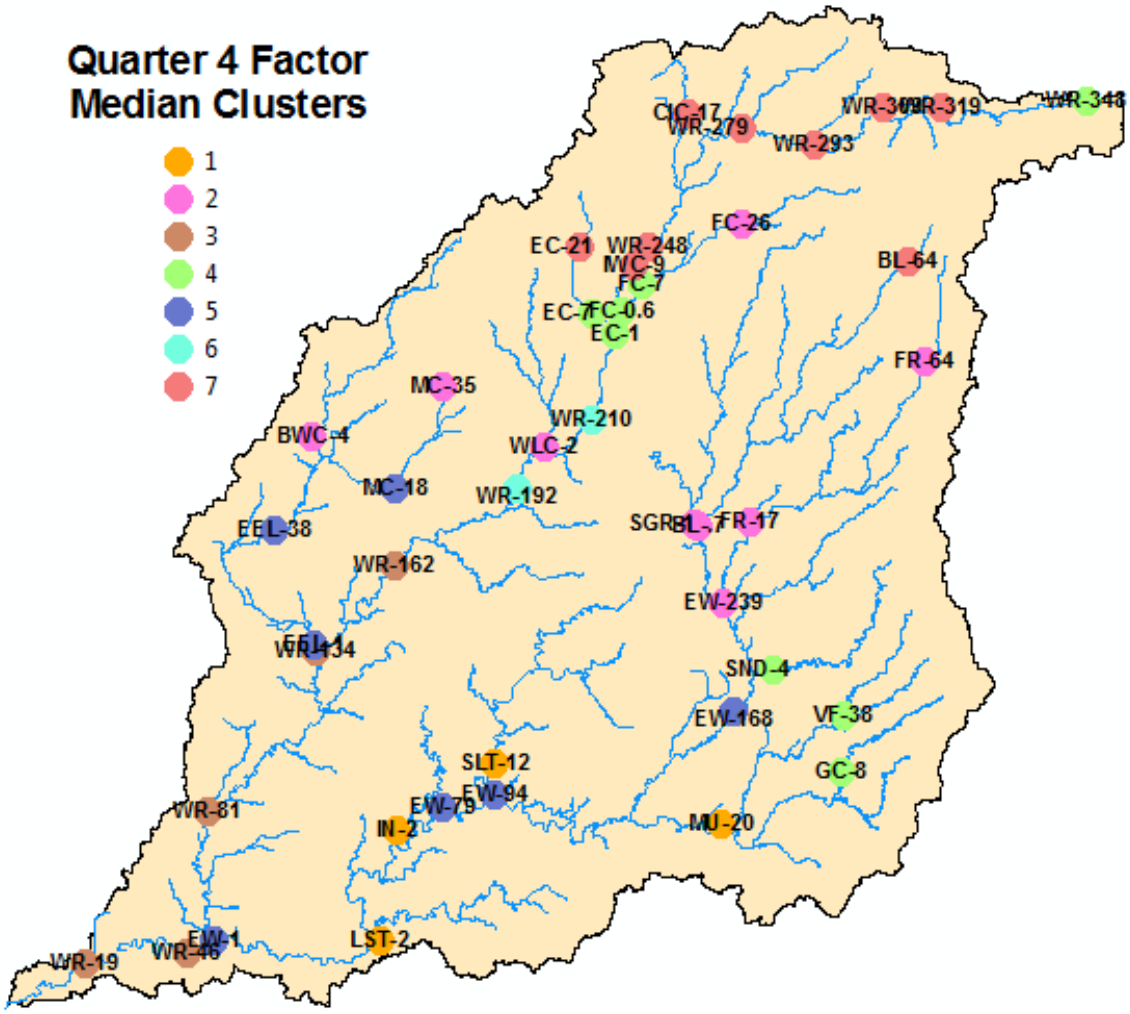
Supplementary Figure 6.31 – Spatially distributed clustering for the Quarter 3 Trimmed Mean SOM (EEL-1 belongs to cluster 3 and WR-134 belongs to cluster 2; SGR-1 and BL-.7 belongs to cluster 4)



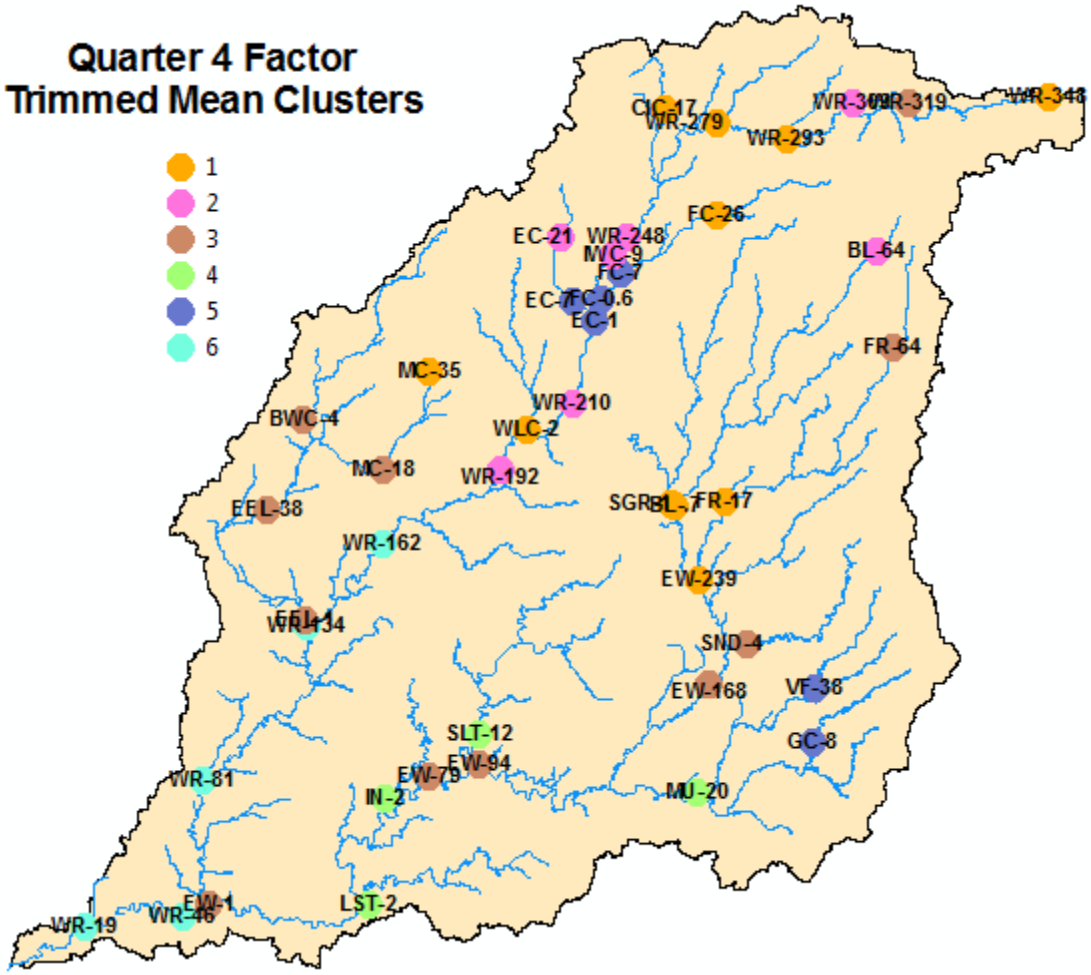
Supplementary Figure 6.32 – Spatially distributed clustering for the Quarter 3 Geometric Mean SOM (EEL-1 belongs to cluster 4 and WR-134 belongs to cluster 1; SGR-1 and BL-.7 belongs to cluster 7)



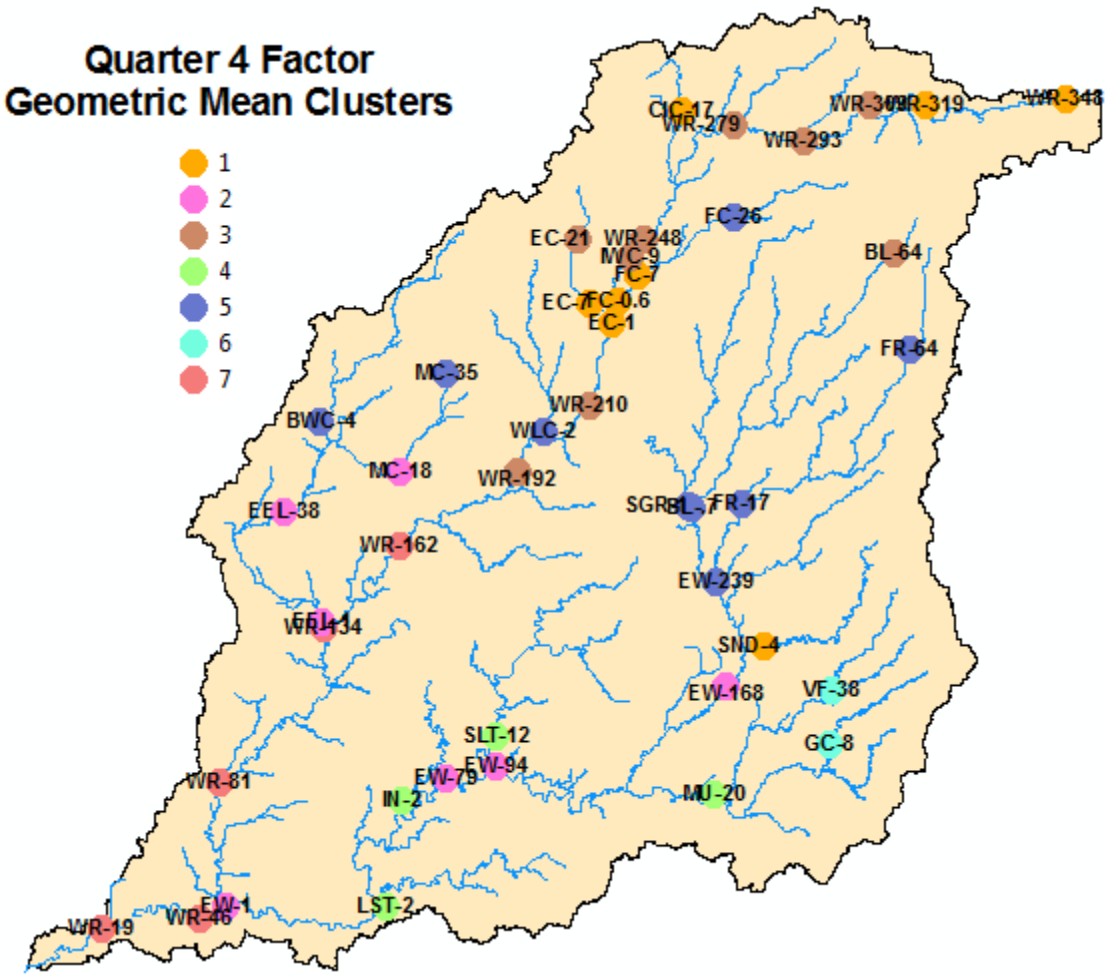
Supplementary Figure 6.33 – Spatially distributed clustering for the Quarter 4 Mean factors (EEL-1 and WR-134 belong to cluster 2; SGR-1 and BL-.7 belongs to cluster 1)



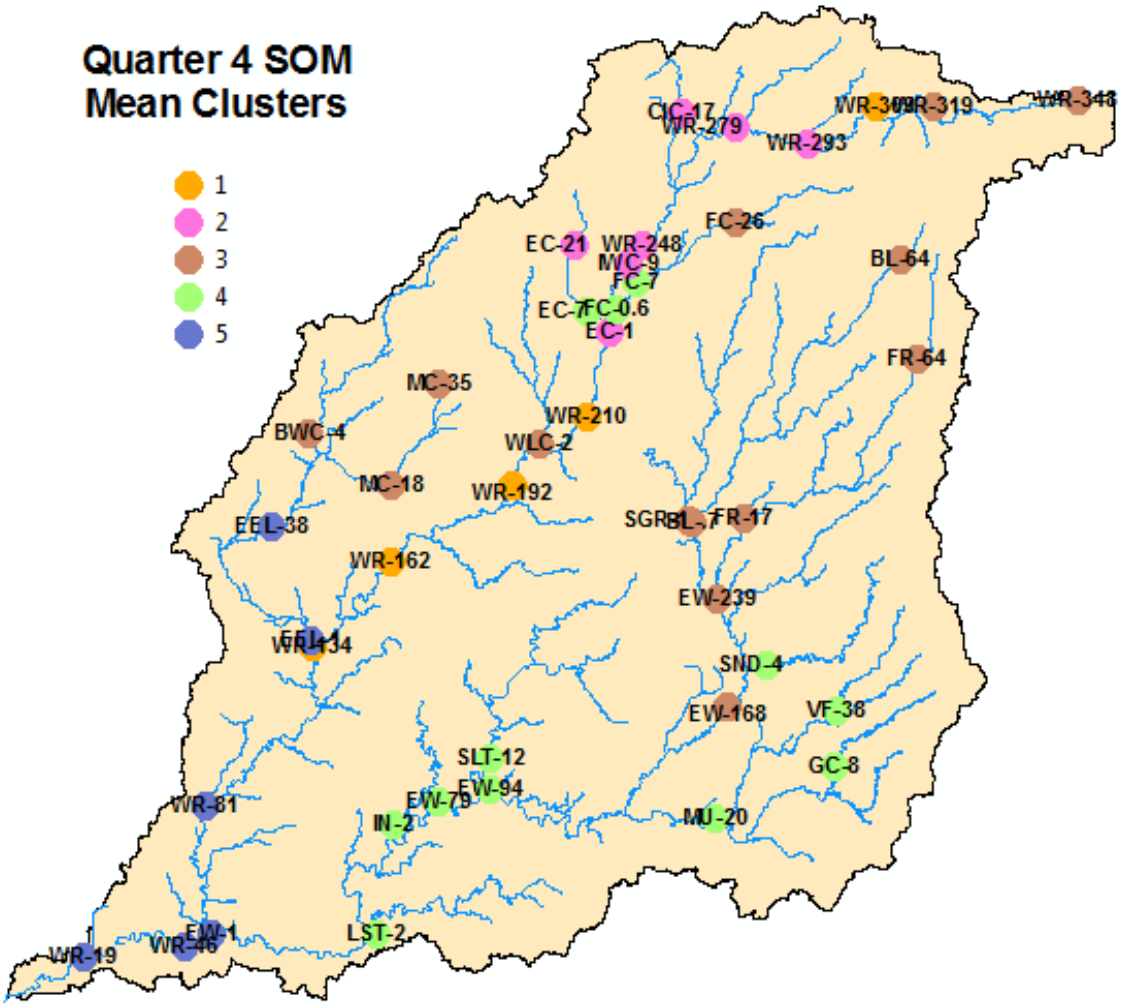
Supplementary Figure 6.34 – Spatially distributed clustering for the Quarter 4 Median factors (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belongs to cluster 2)



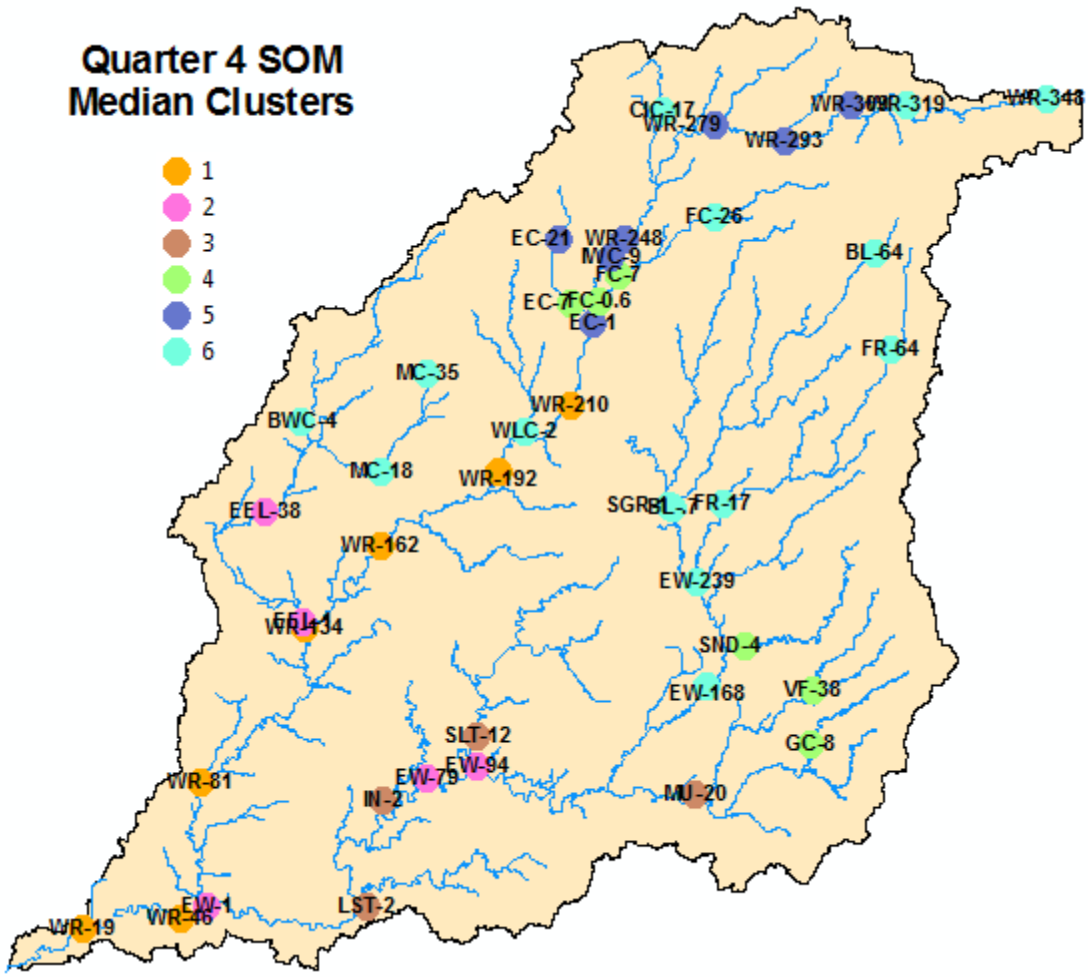
Supplementary Figure 6.35 – Spatially distributed clustering for the Quarter 4 Trimmed Mean factors (EEL-1 belongs to cluster 3 and WR-134 belongs to cluster 6; SGR-1 and BL-.7 belongs to cluster 1)



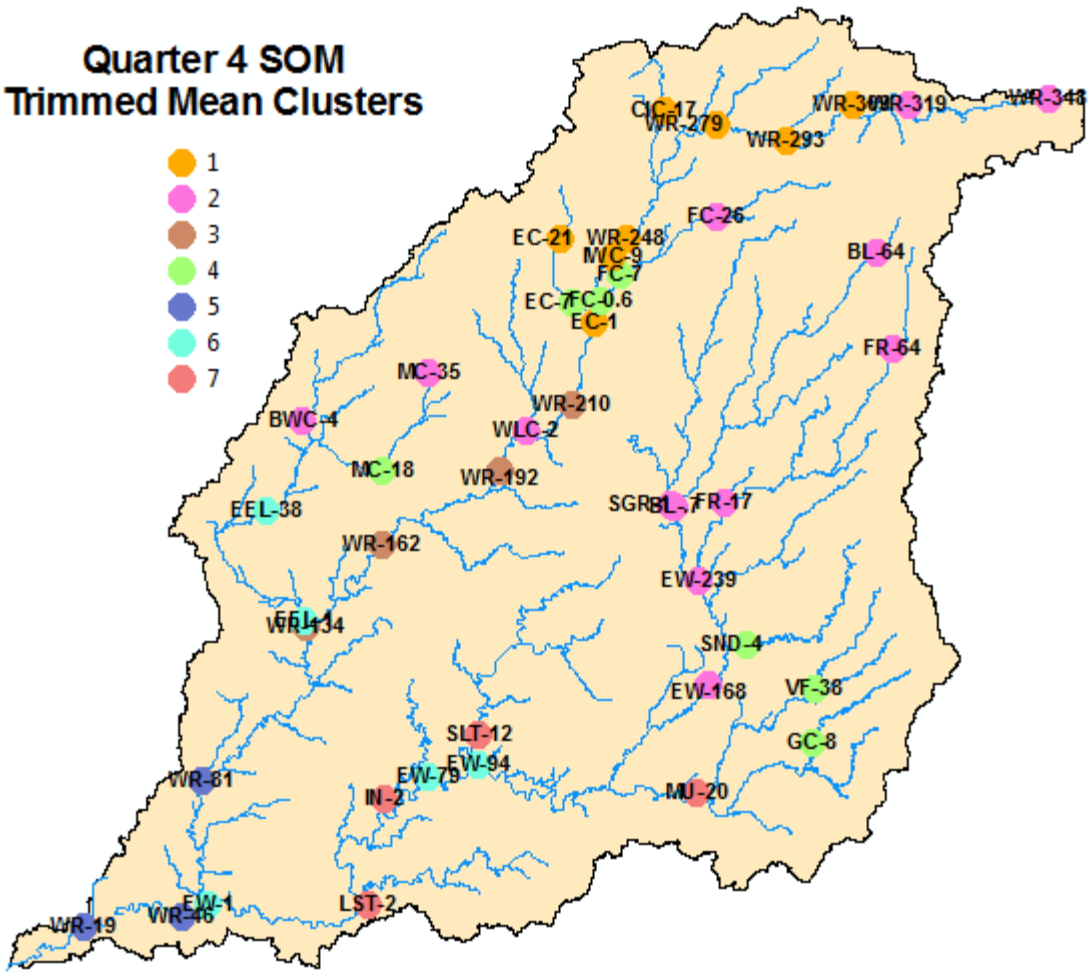
Supplementary Figure 6.36 – Spatially distributed clustering for the Quarter 4 Geometric Mean factors (EEL-1 belongs to cluster 2 and WR-134 belongs to cluster 7; SGR-1 and BL-.7 belongs to cluster 5)



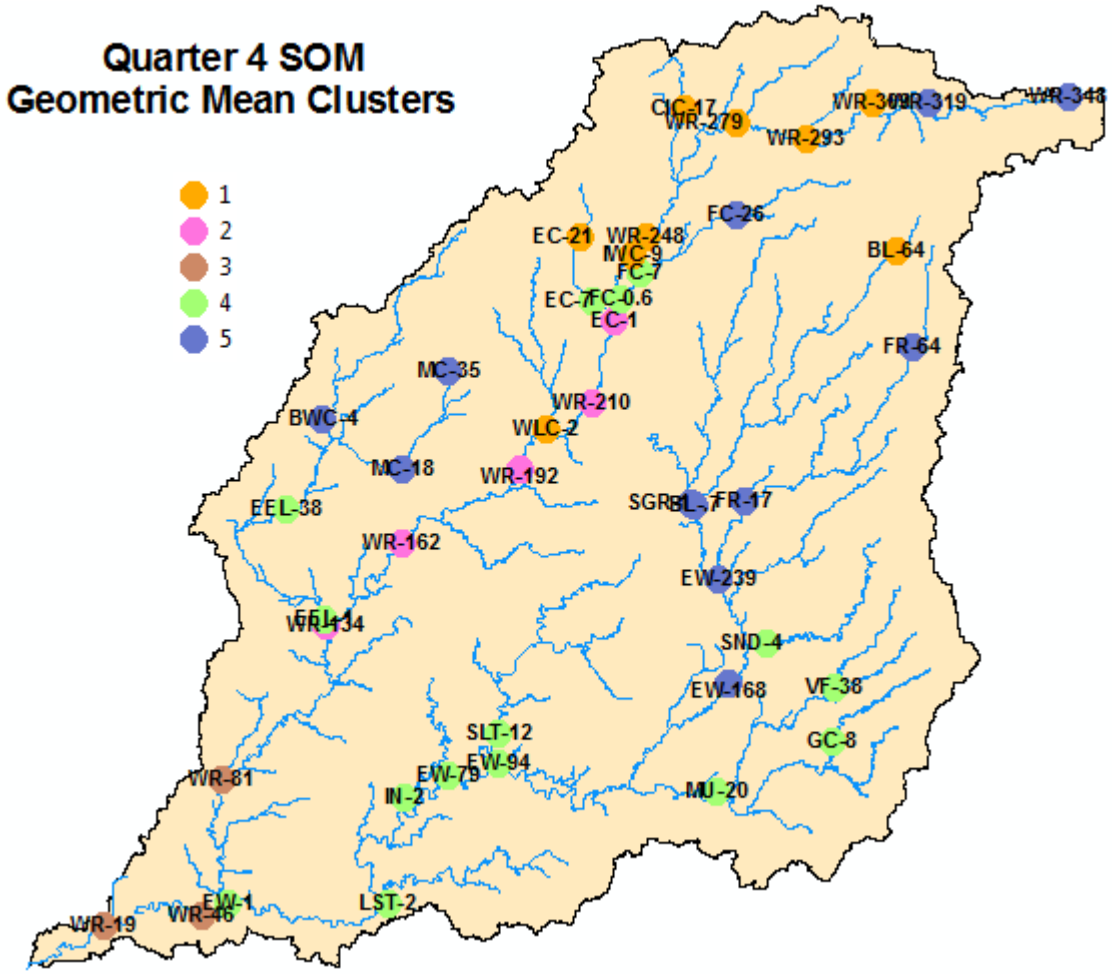
Supplementary Figure 6.37 – Spatially distributed clustering for the Quarter 4 Mean SOM (EEL-1 belongs to cluster 5 and WR-134 belongs to cluster 1; SGR-1 and BL-.7 belongs to cluster 3)



Supplementary Figure 6.38 – Spatially distributed clustering for the Quarter 4 Median SOM (EEL-1 belongs to cluster 2 and WR-134 belongs to cluster 1; SGR-1 and BL-.7 belongs to cluster 6)

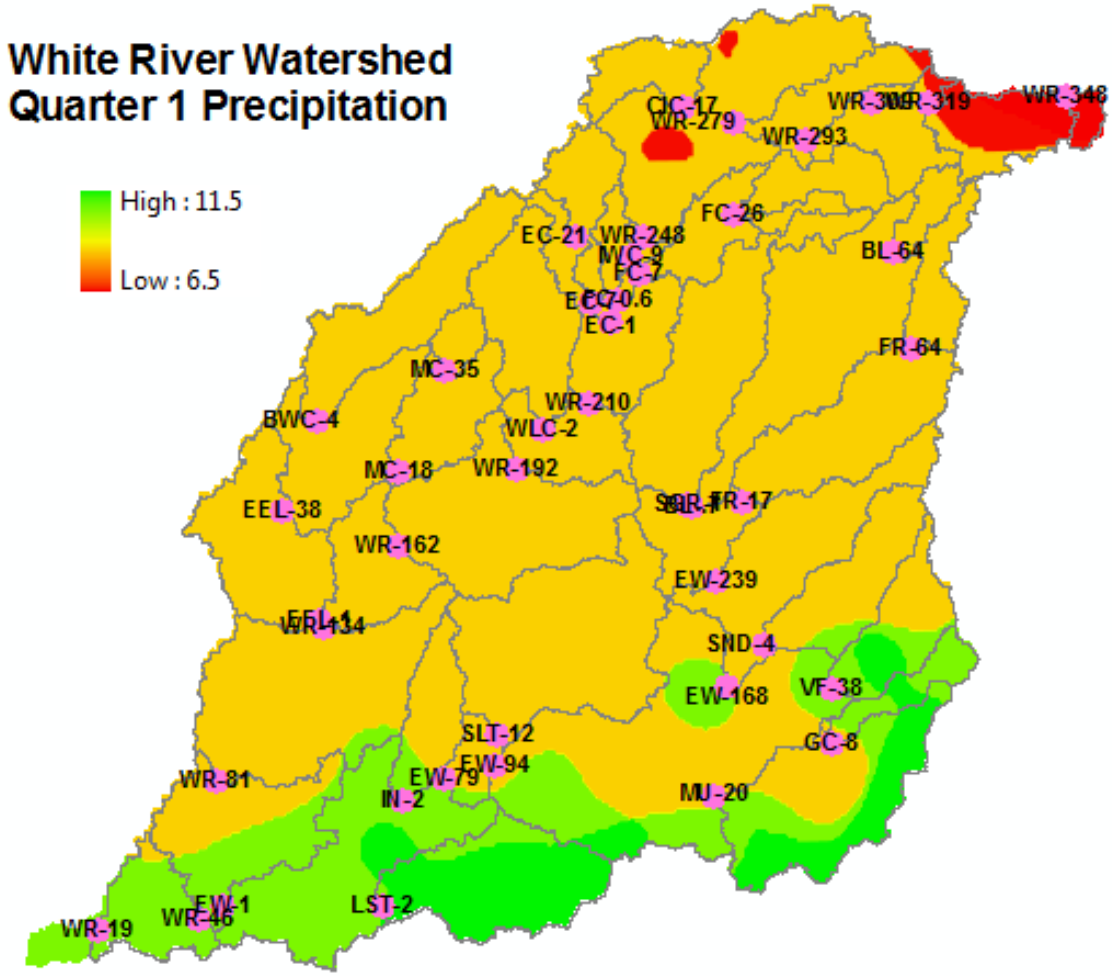


Supplementary Figure 6.39 – Spatially distributed clustering for the Quarter 4 Trimmed Mean SOM (EEL-1 belongs to cluster 6 and WR-134 belongs to cluster 3; SGR-1 and BL-.7 belongs to cluster 2)

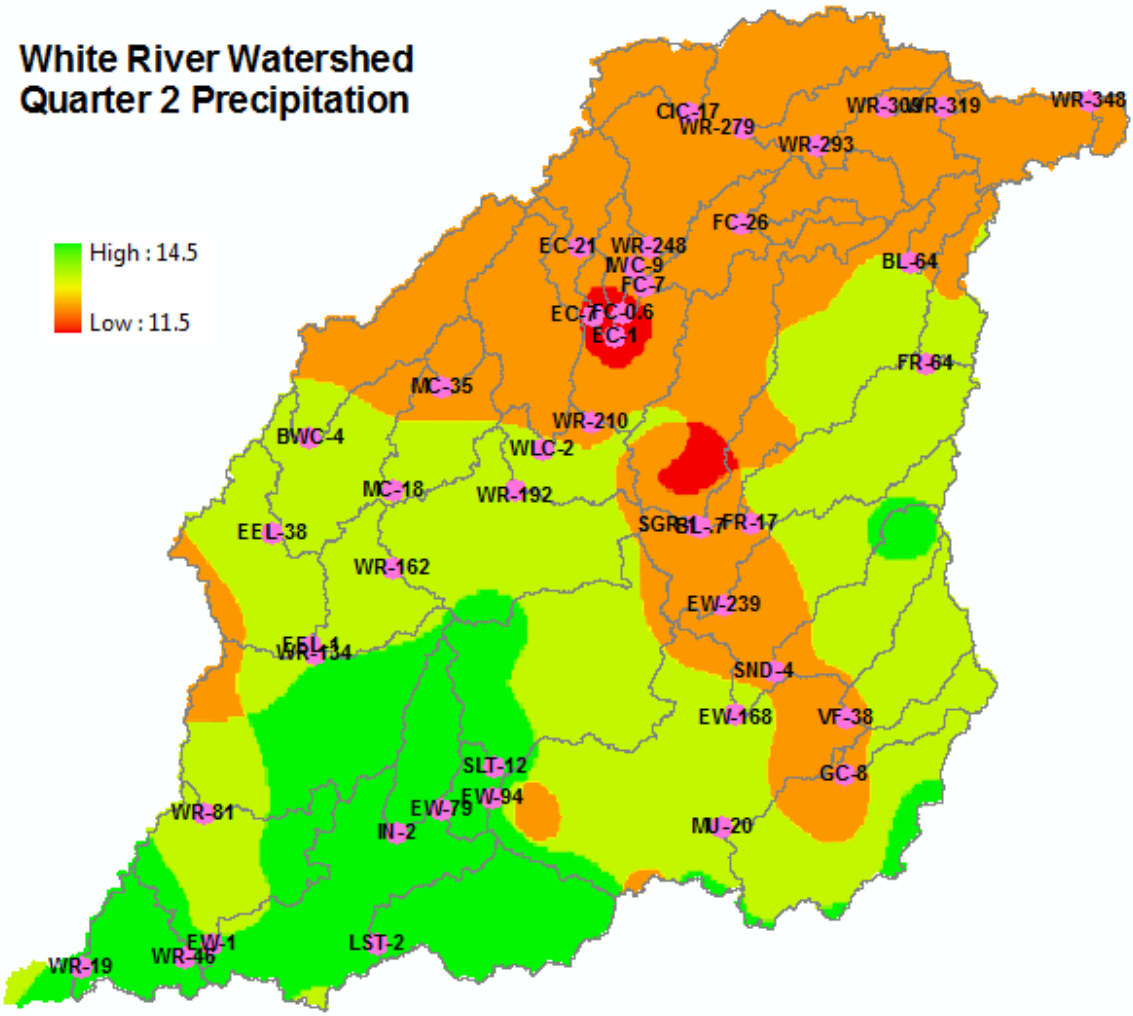


Supplementary Figure 6.40 – Spatially distributed clustering for the Quarter 4 Geometric Mean SOM (EEL-1 belongs to cluster 4 and WR-134 belongs to cluster 2; SGR-1 and BL-.7 belongs to cluster 5)

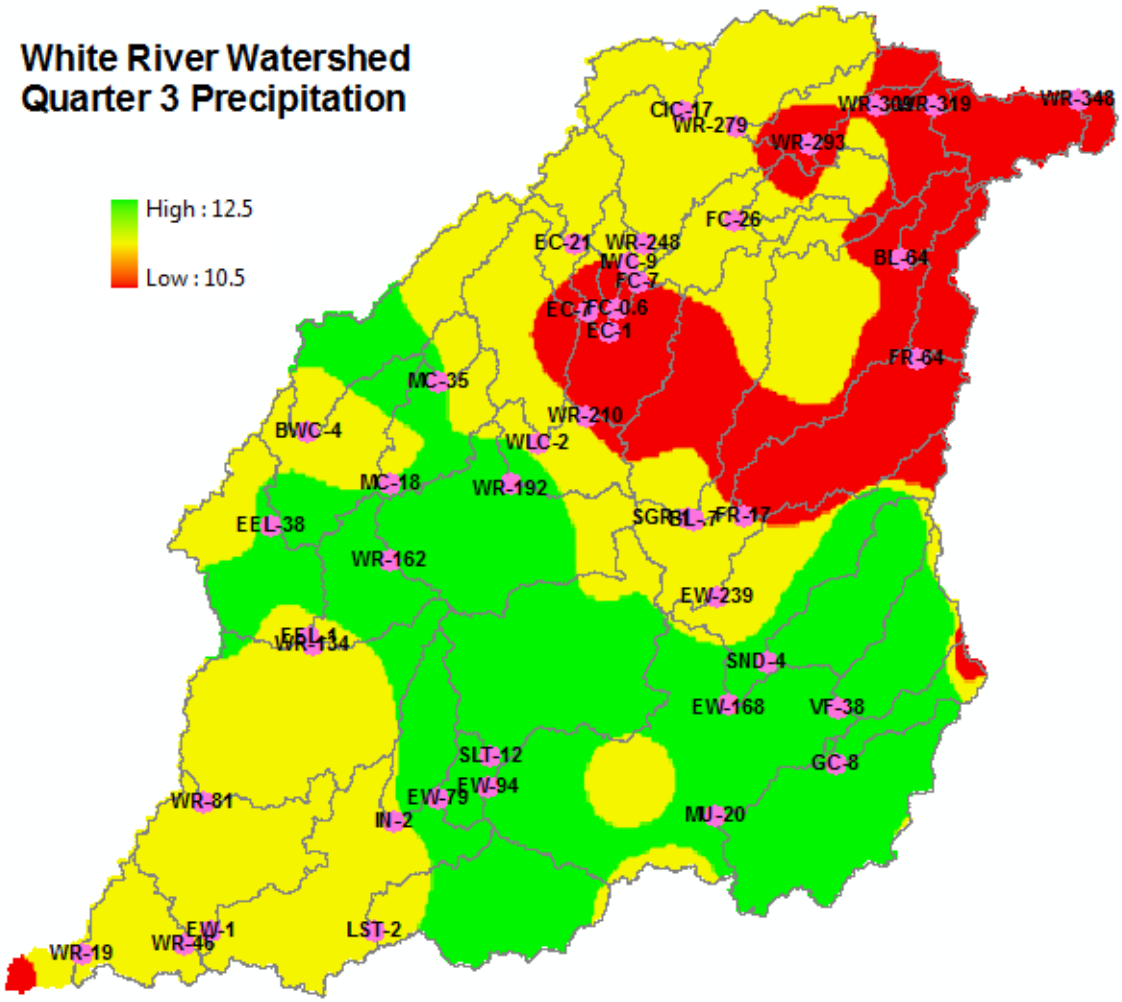
Quarterly White River Watershed Precipitation Maps



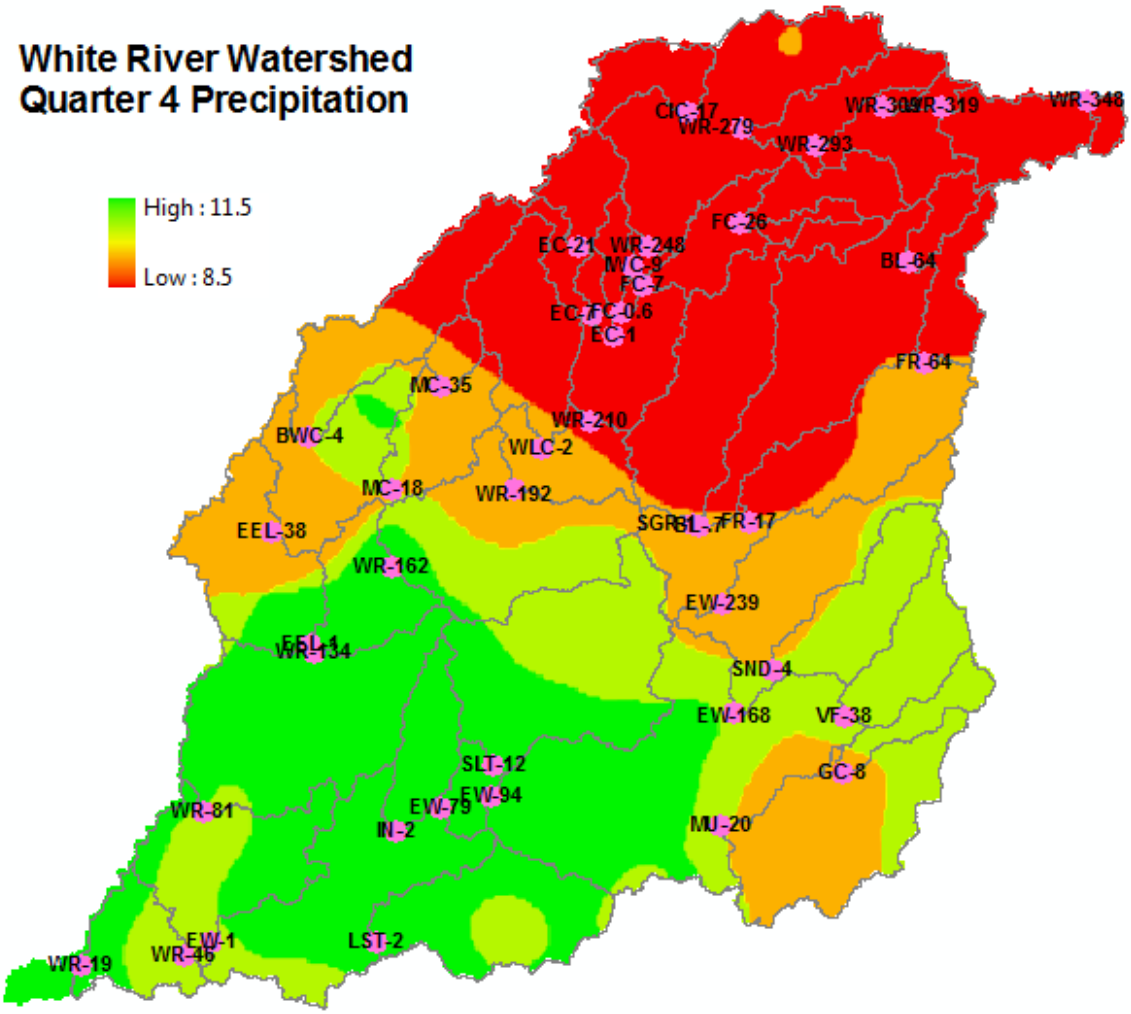
Supplementary Figure 7.1 – White River Watershed mean quarter 1 precipitation (values are in inches)



Supplementary Figure 7.2 – White River Watershed mean quarter 2 precipitation (values are in inches)



Supplementary Figure 7.3 – White River Watershed mean quarter 3 precipitation (values are in inches)



Supplementary Figure 7.4 – White River Watershed mean quarter 4 precipitation (values are in inches)

APPENDIX C – COMMANDS TO RUN STATISTICAL ANALYSES

Principal Component Analysis SAS Code

```
data WQ_PCA;
input alk toc cl cod do hard tkn no2no3 ph P tss sc sul temp turb fe;
datalines; /*Copy and paste raw data without headings or labels.  Be sure the variables
that are copy pasted are in the same order as the input variables*/
0.780113362   0.199904721   0.535835943   0.232059767   0.682869969…
;
/*PCA – Variables were normalized with Box-Cox transformations.  The normalized
variables were then standardized with a softmax transformation that was calculated in the
SOM Toolbox 2.0*/
title 'FA method=prin All Season Geometric Mean';
proc factor data=wq_pca scree rotate=varimax method=prin out=scores nfactors=4
flag=.6 priors=one;
var alk toc cl cod do hard tkn no2no3 ph tss sc sul temp turb fe;
run;
proc print data=scores;
var factor1 factor2 factor3 factor4;
run;
```

Linear Discriminant Analysis SAS Code

```
proc format;
value clusfmt
1='Cluster1'
2='Cluster2'
3='Cluster3'
/*Continue this for however many clusters that there are*/
data WQ ;
title 'Train DA';
/*The ‘station $7.’ indicates that the first column in the data is the station number and the ‘$7’ indicates how many characters this will take up.  The cluster variable at the end is the cluster assignment.  All of the other variables are the softmax transformed physical watershed variables.  Be sure the variables that are copy pasted are in the same order as the listed after ‘input’.*/
input station $7. LFP ND SS Area Slp Temp Precip cornbelt intplat cofos csos npdes is GS LS LSD SSLSShl silt till blue highrim hills water urban for grass crop wet utof utoa ftou ftoa atou atof W MW P SP cluster;
    format Cluster clusfmt.;
    datalines;
BL-.7   0.448139321   0.285024587   0.5345793      0.405556811…
;

/*The code below runs the Stepwise LDA*/
title 'All Geomean';
proc stepdisc data=WQ slentry=0.10 slstay=0.10;
class cluster;
var LFP ND SS Area Slp Temp Precip cornbelt intplat cofos csos npdes is tree GS LS LSD SSLSShl till blue highrim hills water urban for grass crop wet utof utoa ftou ftoa atou atof W MW P SP;
run;

/*The code below runs the parametric LDA*/
proc discrim data=WQ method=normal pool=yes outstat=dis_fun listerr crosslisterr;
class cluster;
var highrim npdes precip urban atou wet utoa sslsshl cornbelt; /*The variables entered here are from the chosen variables from the Stepwise LDA*/
id station;
title2 'All Geomean Using Normal Density Estimates with Equal Variance';
run;
```

```
/*Classify ECWMP data*/
data new;
title 'Test DA';
/*Follows the same preparations that was done for the training input data, except there is no cluster variable*/
input station $9. LFP ND SS Area Slp Temp Precip cornbelt intplat cofos csos npdes is GS LS LSD SSLSShl silt till blue highrim hills water urban for grass crop wet utof utoa ftou ftoa atou atof W MW P SP;
datalines;
ECWMP-01    0.258875512   0.213454398…
;
/*The code below inputs test data into the classification equations*/
proc discrim data=dis_fun testdata=new testlist;
title 'Eagle Creek Test Data';
class Cluster;
/*The variables entered here are from the chosen variables from the Stepwise LDA*/
var highrim npdes precip urban atou wet utoa sslsshl cornbelt;
run;
```

Kohonen Self-Organizing Map MATLAB Commands

```
%       All Kohonen SOMs were built using the source code from the SOM Toolbox 2.0
%       (Alhoniemi, 1999)

%       Labeling different stations.  Data was entered in alphabetical order of the station
%       name, and the stations were in rows.
labels=cell(44,1);
labels{1,1}='BL.7';labels{2,1}='BL-64';labels{3,1}='BWC-4';labels{4,1}='CIC-
17';labels{5,1}='EC-1';labels{6,1}='EC-21';labels{7,1}='EC-7';labels{8,1}='EEL-
1';labels{9,1}='EEL-38';labels{10,1}='EW-1';labels{11,1}='EW-168';labels{12,1}='EW-
239';labels{13,1}='EW-79';labels{14,1}='EW-94';labels{15,1}='FC-
0.6';labels{16,1}='FC-26';labels{17,1}='FC-7';labels{18,1}='FR-17';labels{19,1}='FR-
64';labels{20,1}='GC-8';labels{21,1}='IN-2';labels{22,1}='IWC-9';labels{23,1}='LST-
2';labels{24,1}='MC-18';labels{25,1}='MC-35';labels{26,1}='MU-
20';labels{27,1}='SGR-1';labels{28,1}='SLT-12';labels{29,1}='SND-
4';labels{30,1}='VF-38';labels{31,1}='WLC-2';labels{32,1}='WR-
134';labels{33,1}='WR-162';labels{34,1}='WR-19';labels{35,1}='WR-
192';labels{36,1}='WR-210';labels{37,1}='WR-248';labels{38,1}='WR-
279';labels{39,1}='WR-293';labels{40,1}='WR-309';labels{41,1}='WR-
319';labels{42,1}='WR-348';labels{43,1}='WR-46';labels{44,1}='WR-81';

%       Labeling the variables – the data was organized in this order.  Water quality
%       variables were in columns.
cnames={'alk','TOC','Cl','COD','DO','Hard','TKN','NO2NO3','pH','TotalP','TSS','SC','SO4
','Temp','Turb','Fe'};

%       This command performs the softmax transformation.  All softmax transformations
%       were performed in this toolbox.
sS=som_normalize(data,'logistic');

%       Creates the data structure for SOM construction.
sD=som_data_struct(sS,'labels',labels,'comp_names',cnames);

%       Creates SOM and adds the labels of the the monitoring stations.
sM_data=som_make(sD,'init','randinit','algorithm','seq','mapsize','big','training',[1000
5000]);
sM_data=som_autolabel(sM_data,sD,'add');
```

```
%       Creates the Unified Distance Matrix and Component Map visualizations
som_show(sM_data,'umat','all','empty','Labels');
som_show_add('label',sM_data.labels,'textsize',8,'textcolor','r');
som_show(sM_data)
```

Cluster Analysis MATLAB Commands

```
%       All cluster analyses were done using the source code from the SOM Toolbox 2.0
%       (Alhoniemi, 1999)

%       This command clusters the data.  Both the SOM and factor scores were clustered
%       using this command.  The 'ind' output variable contains the Davies-Bouldin index
%       that helps choose the number of clusters.  The output variable 'p' contains the
%       cluster assignments for different values of k clusters.  A data structure with station
%       labels must be made before the factor scores can be clustered.
[c,p,err,ind]=kmeans_clusters(sM_data or PCAFactors,10,20000);

%       This command creates the SOM Cluster Arrangements.  The variable 'i' tells the
%       program which k cluster arrangement should be used to draw the SOM cluster
%       arrangements.  For example, for 5 clusters the command would be
%       'som_show(sM_data,'color',{p{5},sprintf('%d clusters',5)});'.
som_show(sM_data,'color',{p{i},sprintf('%d clusters',i)});
som_show_add('label',sM_data.labels,'textsize',8,'textcolor','w');
```

Support Vector Machine MATLAB Commands

```
%       Use the softmax transformation in the SOM Toolbox 2.0 to preprocess the
%       physical watershed parameters.
%       All SVMs were built using LibSVM – A Library for Support Vector Machines in
%       a simple MATLAB interface (Chen and Lin, 2001).

%       These commands train the SVM using grid-search and leave-one-out cross
%       validation.  Use the cluster assignments as the targets and the physical watershed
%       parameters as the training input.  Choose hyperparameters [C g] from best cross
%       validation (record this value).  Choose lower C parameter if the same cross
%       validation accuracy occurs twice.  This will provide a better generalization.
bestcv=0;
for log2c = -1:10,
for log2g = -4:1,
cmd = ['-t 2 -v 44 -c ', num2str(2^log2c), ' -g ', num2str(2^log2g)];
cv = svmtrain(Targets, Training_Input, cmd);
if (cv >= bestcv),
bestcv = cv; bestc = 2^log2c; bestg = 2^log2g;
end
fprintf('%g %g %g (best c=%g, g=%g, rate=%g)\n', log2c, log2g, cv, bestc, bestg,
bestcv);
end
end

%       Create model using the chosen hyperparameter values and create a model to save
%       for validating data later.
model = svmtrain(Targets, Training_Input, '-t 2 -c ? -g ? -b 1');

%       Validate model with the ECWMP stations' physical watersheds parameters.  The
%       target values do not matter as long as there is a value for each station.
[predict_label, accuracy, prob_est] = svmpredict(Targers, Test_Input, model, '-b 1');
```

ProUCL 4.0 Instructions

Regression on Order Statistical (ROS) Method:
This method was used to estimate non-detectable data before time-averaging took place.

1. To begin arrange the water quality variable including non-detectable data in a column in the ProUCL 4.0. Leave the data above the detection limit as is, and enter the detection limit value in each cell where the data point was below the detection limit. For example, a datapoint below a 0.2 mg/L detection limit may be labeled "<0.2". In this case, 0.2 would be entered into that cell.

2. Make a new column to the left of the dataset of interest. For each adjacent cell, indicate if the data is above the detection limit with a 1. If the data is non-detectable data, indicate this with a 0. The first to steps may be done in Excel for ease of organizing and manipulating the data. However, beginning with Step 3 the data should be in the ProUCL 4.0 application.

3. Once the data is arranged properly, label the column with the raw water quality data "data", and label the adjacent column "D_data" by right clicking on the column header.

4. Click on the "ROS Est. NDs" menu and select "Lognormal ROS".

5. In the "Select Variables" window click the given name of the dataset of interest, then click the right arrows. The data to be estimated should be in the "Selected" section.

6. Click "OK". A third column will appear labeled "LnROS_data". This column contains the ROS estimates of the non-detectable data.

7. Copy this column of data into the original spreadsheet and proceed with calculating each statistical indicator.

REFERENCES

Akume, D. and Weber, G.-W. 2002. Cluster Algorithms: Theory and Methods. *Journal of Computational Technologies*, 7(1):15-27.

Alhoniemi, E., Himberg, J., Parviainen, J., Vestano, J. 1999. SOM Toolbox 2.0, a software library for Matlab 5 implementing the Self-Organizing Map algorithm. Retreived from http://www.cis.hut.fi/somtoobox.

Bezdek, J.C. and Pal, N. R. 1998. Some New Indexes of Cluster Validity. IEEE transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 28(3):301-315.

Bowers, J.A., Shedrow, C.B. 2000. Predicting Stream Water Quality Using Artificial Neural Networks (ANN). Development and Application of Computer Techniques to Environmental Studies VIII. WIT Press Southampton, England, UK. pgs. 89-98.

Box, G.E.P. and Cox, D.R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological),* 26(2):211-252.

Chang, C., Lin, C. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, Y., Lin, C. 2006. Combining SVMs with Various Feature Selection Strategies. Studies in Fuzziness and Soft Computing, *Springer Berlin/Heidelberg*. Vol. 207, pgs, 315-324.

Collica, R.S. 2007. CRM Segmentation and Clustering Using SAS Enterprise Miner. SAS Institute Inc., Cary, NC. pgs. 100-101.

Costanza, M.C. and Afifi, A.A. 1979. Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis. *Journal of the American Statistical Association,* 74:777-785.

Davis, J.C., 2002. Statistics and Data Analysis in Geology (Third Edition). John Wiley and Sons, Inc., New York, NY.

Dalzell, B.J., Filley, T.R., Harbor, J.M. 2006. The role of hydrology in annual organic carbon loads and terrestrial organic matter export from a Midwestern agricultural watershed. *Geochica et Cosmochimica Acta,* 71:1448-1462.

ESRI. 2005. Arc Hydro – HydroID. Version 1.1 Final, July 2005.

Fodor, I. 2002. A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA.

Frohlich, H.L., Breuer, L., Frede, H., Huisman, J.A., Vache, K.B. 2007. Water Resource characterization through spatiotemporal patterns of major, minor and trace element stream concentrations in a complex, mesoscale German catchment. *Hydrological Processes*, 22:2028-2043.

Fry, J.A., Coan, M.J., Homer, C.G., Meyer, D. K., Wickham, J.D. 2009. Completion of the National Land Cover Database (NLCD) 1992-2001 Land Cover Change Retrofit product. USGS Open-File Report 2008-1379, 18 p.

Fenelon, J.M. 1998. Water Quality in the White River Basin, Indiana, 1992-1996: U.S. Geological Survey Circular 1150.

Gunn, S.R. 1998. Support Vector Machines for Classification and Regression. Technical Report, University of Southhampton.

Hammer, O., Harper, D.A.T., Ryan, P.D. 2009. PAST – Palaeontological STastitics, ver. 1.89. Technical Report.

Hellweger, F. 1997. AGREE – DEM Surface Reconditioning System. University of Texas at Austin, http://www.ce.utexas.edu/prof/maidment/GISHYDRO/ferdi/research/agree/agreht#Part2

Homer, C.C., Huang, C., Yang, L., Wylie, B., Coan, M. 2004. Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing*, 70(7):829-840.

Hem, J., 1985. Study and Interpretation of the Chemical Characteristics of Natural Water (Third Edition). US Geological Survey Water Supply Paper 2254.

Hsu, C., Change, C., Lin, C. 2010. A Practical Guide to Support Vector Classification. Technical report, Dept. of Computer Science, National Taiwan University, Taipei 106, Taiwan. http://www.csie.ntu.edu.tw/~cjlin.

Iscen, C.F., Emiroglu, O., Arslan, N., Yilmaz, V., Ahiska, S. 2008. Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. *Environmental Montitoring Assessment*, 144:269-276.

Jacques, D.V., and Crawford, C.G., 1991, National Water Quality Assessment Program White River Basin: U.S. Geological Survey Open-File Report 91-169, 2 p. (WATER FACT SHEET).

Jenerette, G.D., Lee, J., Waller, D.W., Carlson, R.E. 2002. Multivariate Analysis of the Ecoregion Delineation for Aquatic Systems. *Environmental Management,* 29:67-75.

Jiang, Y. and Nan, Z. 2006. Integration of Artificial Neural Network with GIS in Uncertain Model of River Water Quality. National Laboratory of Western China's Environmental Systems and College of Resource and Environment Sciences, Lanzhou University, Lanzhou, China.

Kartoun, U., Stern, H., Edan, Y. 2006. Bag Classification Using Support Vector Machines. In *Applied Soft Computing Technologies: The Challenge of Complexity*. pgs. 665-674.

Kecman, V. 2001. Learning and Soft Computing – Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Slides accompanying book). The MIT Press, Cambridge, MA.

Kutner, M., Nachtsheim, C., Neter, J. 2004. Applied Linear Regression Models (Fourth Edition). The McGraw Hill Companies, Inc., New York, NY.

Nilsson, R., Pena, J.M., Bjorkegren, J., Tegner, J. 2006. Evaluating feature selection for SVMs in high dimensions. *Proceedings of the 17th European conference on machine learning,* 719-726.

Park, Y. 2003. Deliverable 12: Publication of ANN Model Results. PAEQANN, European Commission, Contract No. EVK1-CT1999-00026. *Available at http://aquaeco.ups-tlse.fr/.*

Paul, S., Srinivasan, R., Sanabria, J., Haan, P.K., Mukhtar, S., Neimann, K. 2006. Groupwise Modeling and Study of Bacterially Impaired Watersheds in Texas: Clustering Analysis. *Journal of the American Water Resources Association*, Paper No. 04216.

Rao, A.R., Srinivas, V.V. 2008. Regionalization of Watersheds. Springer Science+Business Media B.V. Water and Science Library of Technology. Volume 58.

Ren, Y., Liu, H., Xue, C., Yao, X., Liu, M., Fan, B. 2006. Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. *Analytica Chimica Acta,* 572:272-282.

Rojas, R. 1996. Neural Networks – A Systematic Introduction. Springer-Verlag, Berlin. pgs. 391-412.

Santos-Roman, D.M., Warner, G.S., Scatena, F. 2003. Multivariate Analysis of Water Quality and Physical Characteristics of Selected Watersheds in Puerto Rico. *Journal of the American Water Resources Association*, Paper No. 01039.

SAS (SAS Institute Inc.). 2002-2004. SAS 9.1.3 Help and Documentation. SAS Institute, Inc., Cary, NC.

Siegel, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company, New York, NY.

Singh, A., Maichle, R., and Lee, S. 2006. On the Computation of a 95% Upper Confidence Limit of the Unkown Population Mean Upon Data Sets with Below Detection Limit Observations. USEPA, Contract No. 68-W-04-005, Las Vegas, NV.

Singh, A., Maichle, R., Singh, A.K., Lee, S.E., Armyba, N. 2007. ProUCL Version 4.00.02 User Guide. US EPA, Office of Research and Development. Washington, DC.

Soltani, S., Modarres, R. 2006. Classification of Spatio-Temporal Pattern of Rainfall in Iran Using A Heirarchical and Divisive Cluster Analysis. *Journal of Spatial Hydrology*, 6(2).

Suhr, D.D. 2005. Principal Component Analysis vs. Factor Analysis. SAS SUGI 30 Proceedings, Statistics and Data Analysis Section, Paper No. 203-30, SAS Institute Inc., Cary, NC.

Snelder, T.H., Biggs, B.J.F., Woods, R.A. 2005. Improved Eco-hydrological Classification of Rivers. *River Research and Applications,* 21:609-628.

Tabachnick, B.G. and Fidell, L.S. 1989. Using Multivariate Statistics (2nd Edition). Harper and Row, Publishers, New York, NY.

Tang, Y., Zhang, Y., Chawla, N.V., Krasser, S. 2002. SVMs Modeling for Highly Imbalanced Classification. *Journal of Latex Class Files*, 1(11).

Tedesco, L.P., Pascual, D.L., Shrake, L.K., Casey, L.R., Vidon, P.G.F., Hernly, F.V., Salazar, K.A., Barr, R.C., Ulmer, J., Perching, D. 2005. Eagle Creek Watershed Management Plan: An Integrated Approach to Improved Water Quality. Eagle Creek Watershed Alliance, CEES Publication 2005-2007, IUPUI, Indianapolis, IN.

USDA. 2004. State Soil Geographic (STATSGO) Data Base – Data use information. Natural Resources Conservation Service, National Soil Survey Center. Miscellaneous Publication Number 1492.

USEPA. 1996. U.S. EPA NPDES Permit Writers' Manual. Office of Water; EPA-833 B-96-003.

USEPA. 2007. An Introduction to Water Quality Monitoring. *Available at http://www.epa.gov/owow/monitoring/monintr.html. Accessed on April 24, 2009. Last updated on March 21st, 2007.*

Vesanto, J. and Alhoniemi, E. 2000. Cluster of the Self Organizing Map. *IEEE Transactions on Neural Networks.* 11(3):586-600.

Vestano, J., Himberg, J., Alhoniemi, E., Parhankangas, J. 2000. SOM Toolbox for Matlab 5. SOM Toolbox Team. Helsinki University of Technology. Report A57

Ward, A. and Trimble, S., 2004. Environmental Hydrology (Second Edition). Lewis Publishers, Boca Raton, Florida.

Woods, A.J., Omerik, J.M., Brockman, C.S., Gerber, T.D., Hosteter, W.D., and Azevedo, S.H. 1998. Ecoregions of Indiana and Ohio (color poster with map, descriptive text, summary tables, and photographs). U.S. Geological Survey, Reston, VA.

Yunrong, X., Liangzhong, J. 2009. Water Quality Prediction Using LS-SVM and Particle Swarm Optimization. Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining, pgs. 900-904.

CURRICULUM VITAE

Andrew Stephan Gamble

Education:

8/08 – 9/10 MS, Department of Earth Sciences, Indiana University – Purdue University Indianapolis

Thesis: COMBINING MULTIVARIATE STATISTICAL METHODS AND SPATIAL ANALYSIS TO CHARACTERIZE WATER QUALITY CONDITIONS IN THE WHITE RIVER BASIN, INDIANA, U.S.A.

8/02 – 5/06 BA, Business Administration, Hanover College

Professional Experience:

8/08 – 9/10 Research Assistant/Graduate Student, Indiana University – Purdue University of Indianapolis

2/07 – 11/07 Crew Leader, Montana Conservation Corps

Conferences Presentations and Proceedings:

Gamble, A. and Babbar-Sebens, M., "Combining Multivariate Statistics and GIS to Characterize Water Quality Conditions in the White River", Central Indiana Water Resources Partnership Science Meeting, Indianapolis, IN, May 10th 2010.

Gamble, A. and Babbar-Sebens, M., "Combining Spatial Analysis and Multivariate Statistical Methods to Characterize Watershed Water Quality Conditions", Proceedings of the AWRA Specialty Conference GIS and Water Resources VI in Orlando, FL March 29 -31, 2010.

Publications:

Gamble, A., and Babbar-Sebens, M. 2010. Combining Multivariate Statistical Methods and Spatial Analysis to Characterize Water Quality Conditions in the White River. To be submitted.

Honors, Awards, Fellowships:

2010 IUPUI Graduate Student Organization Travel Grant Award